A STUDY

IN THE VALIDATION

OF PROFICIENCY TESTS OF

ENGLISH AS A FOREIGN LANGUAGE

ALAN D MOLLER

PhD
University of Edinburgh
1981

Acknowledgements

This study could not have been completed without the cooperation of a large number of people, most of whom I have never met. I am most grateful to all the university and other tutors who completed and returned ratings forms, to the one hundred and six students from overseas who volunteered to undergo up to ninety minutes of language testing, and to the staff in the British Council's Home Division in London and in their regional offices who helped me to contact students and tutors and to obtain data.

I am particularly appreciative of the support and understanding I have received from my family and from my colleagues in the English Language Division of the British Council in London. Above all I wish to express my gratitude to Dr Alan Davies without whose encouragement and guidance this study would never have been completed.

This thesis represents

my own work and has been

composed by myself.

Alan D Moller

ABSTRACT

After examining the development of English language proficiency tests in the light of changes in linguistic theory and psychometric practice, some studies in test validation were reviewed and a lack of information on test validity noted. To meet the need for more information an approach by way of the continuous validation of proficiency tests was proposed, based on internal consistency and on the periodic examination of the performance of different samples on appropriate criterion measures.

To test this approach two investigations into the validity of the English Proficiency Test Battery - EPTB - (Davies, 1964) and the British Council Subjective Assessment - BCSA - were carried out. Two criterion measures were developed, an English Ability Rating form and the Communicative Proficiency Measure, and two samples of students recently arrived in Britain whose English had previously been assessed by EPTB or BCSA were identified (total N = c 1,800). The students' tutors completed the English Ability Ratings for both samples, and the Communicative Proficiency Measure was administered to over 10% of the subjects in the second investigation.

Findings showed that the criterion measures possessed an acceptable degreee of internal and concurrent validity, that EPTB was consistent with the underlying theory and BCSA was consistent with traditional testing practice. Strong positive relationships were established between subjects' performance on either EPTB or BCSA and the criterion measures. Further

analysis of results indicated that the following of pre-sessional English courses affected the relationship adversely, although the relationship remained strong for subjects following the more academic courses and whose English had been assessed by EPTB.

It was concluded that the proposed approach to test validation was practicable and yielded appropriate information on the internal validity and predictive validity of two different types of proficiency test. Studies of new language proficiency tests could appropriately use a similar approach.

TABLE OF CONTENTS

		Page
CHAPTER 1	ENGLISH LANGUAGE PROFICIENCY TESTING	
1.0	Language Examining and Testing - background	1
1.1	Traditional Test Types	2
1.1.1	Their aims	3
1.1.2	Their characteristics	4
1.1.3	Introduction of proficiency testing	4
1.2	Post-War Developments in Language Testing	5
1.2.1	Influence of the work of Lado and Carroll	6
1.2.2	Characteristics of tests of this period	8
1.2.3	Components of a typical test	9
1.2.4	Analysis of six proficiency tests	10
1.2.5	Test CGM 62	12
1.2.6	CPE	13
1.3	Current Developments in Language Testing	14
1.3.1	Characteristics of tests of the current period	15
1.3.2	Notion of overall proficiency	16
1.3.2.1	The white noise test	17
1.3.2.2	Dictation tests	19
1.3.2.3	Cloze technique	20
1.3.3	Inadequacies of these tests	22
1.3.4	FSI proficiency tests	23
1.3.5	Problem of specifications	25
1.3.6	Some tests of functional English ability	25
1.3.7	Characteristics of such tests	26

		Page
1.4	Summary	27
1.4.1	Examiner based testing	27
1.4.2	Language based testing	27
1.4.3	Proficiency based testing	27
1.4.4	Function based testing	28
1.4.5	Differences between the types	28
1.4.6	Conclusion	29

CHAPTER 2 THE PROBLEM OF VALIDITY

2.0	Concept of Validity	31
2.1	Types of Validity	32
2.1.1	Criterion-related validity	32
2.1.2	Content validity and sampling	32
2.1.3	Difficulties of proving validity in language testing	33
2.2	Validity Studies of Language Proficiency Tests	34
2.2.1	Lack of emphasis on validity	35
2.2.2	Growing concern among test users	36
2.3	Content Validity	36
2.3.1	Problem of defining language proficiency	37
2.3.2	Problem of sampling	39
2.3.3	Purpose of the language proficiency	39
2.3.4	Content validity as a continuum	40
2.4	Construct validity	40
2.4.1	The construct of proficiency	41

		Page
2.4.2	Need for proficiency model	42
2.4.3	Need for tests of production	43
2.4.4	Construct validity study - TOEFL	44
2.4.5	Sub-test intercorrelations	45
2.4.6	Correlations with other measures	46
2.4.7	Context and language in the construct of proficiency	47
2.5	Criterion Validity	48
2.5.1	Concurrent and predictive criteria - differences	48
2.5.2	Concurrent validity criteria	50
2.5.2.1	Teachers' ratings	51
2.5.2.2	Other proficiency tests	53
2.5.2.3	Specially constructed measures - the Pike study	53
2.5.3	Predictive validity criteria	55
2.5.3.1	Nature of the criterion measures	56
2.5.3.2	Grade point average	56
2.5.3.3	Examination results	58
2.5.3.4	Academic success	59
2.5.3.5	Special Achievement Index	60
2.5.3.6	Other achievement criteria	61
2.5.3.7	Heaton and Pugh's study	62
2.5.3.8	Gue and Holdaway's study	64
2.5.3.9	Sharon's study	65
2.5.3.10	Language proficiency and academic success	65
2.5.4	The validity of the criteria summarised	66
2.5.5	Use of correlations and expectancy tables	67
2.6	Conclusions on the Validity of Proficiency Tests	68
2.6.1	No certainty of validity	68

		Page
2.7	The Test Consumer's Viewpoint	69
2.7.1	Intuitive judgements and follow-up data	69
2.7.2	Student numbers	70
2.7.3	Need for fresh look at test validity	71
2.8	An Approach to Test Validation	72
2.8.1	Consistency of construction	72
2.8.2	Acceptability of performance	73
2.8.2.1	Reliability	73
2.8.3	The approach summarised	74
2.8.3.1	Nature of information obtained	74
2.8.3.2	An on-going process	75
2.8.4	Establishing the validity of the approach	76
2.8.5	The need to investigate the approach	76
2.8.6	Measures chosen for the investigation	76
2.9	Formulation of Hypotheses	77
2.9.1	First hypothesis	77
2.9.2	Second hypothesis	78
2.9.3	Third hypothesis	78
2.9.4	Fourth hypothesis	78
2.9.5	Fifth hypothesis	78
2.10	Stages of the Investigation	78
CHAPTER 3	ENGLISH PROFICIENCY MEASURES INVESTIGATED	
3.0	Introduction	80

		Page
3.1	Design of EPTB	80
3.1.1	Linguistic content	80
3.1.2	Work sample subtests	81
3.1.3	Try-out	81
3.1.4	Short Version and parallel forms	81
3.1.5	Scoring	83
3.1.6	Description of subtests	84
3.1.6.1	Test of phonemic discrimination	84
3.1.6.2	Test of intonation	84
3.1.6.3	Test of grammar	85
3.1.6.4	Test of reading comprehension	86
3.1.6.5	Test of reading speed	87
3.1.7	Reliability	87
3.1.8	Validity claimed	88
3.2	Consistency of EPTB with the Theory	88
3.2.1	Basic theory of the test	89
3.2.2	The linguistic level tests	89
3.2.3	Work sample tests	91
3.2.4	Validity	94
3.2.5	Consistency of the Battery	95
3.2.5.1	'Linguistic' and 'language control' balance	95
3.2.6	Current concern with communication	96
3.2.7	EPTB and current thinking	97
3.3	Design of BCSA	98
3.3.1	Historical development	98
3.3.2	Basis of the test	99
3.3.3	Components of the test	100

		Page
3.3.4	Responsibility of the assessor	103
3.3.5	Reliability and validity	104
3.4	Consistency of BCSA with the Theory	105
3.4.1	Basic principles	106
3.4.2	BCSA and current thinking	106
3.5	Conclusions	107
CHAPTER 4	THE SAMPLE AND THE CRITERION MEASURES	
4.0	Preparatory Phase of the Enquiry	109
4.1	The Sample: Considerations and Constraints	109
4.1.1	Size	110
4.1.1.1	Type of sample	110
4.1.1.2	Range of ability represented	111
4.1.2	Availability	111
4.1.2.1	Constraints	112
4.1.2.2	Cooperation of subjects and tutors	112
4.1.3	Length of stay	113
4.1.4	Consideration of other factors	115
4.1.5	Summary of factors governing the sample	115
4.2	The Criterion: Considerations	116
4.2.1	Academic success as criterion	116
4.2.2	Adequacy of English as a criterion	118
4.2.2.1	Absence or presence of adequacy	119
4.2.3	Adequacy of English chosen as criterion	119

		Page
4.3	Sources of information considered	119
4.3.1	Student self-assessment	120
4.3.2	Tutors' assessment	121
4.3.3	Language tests	122
4.3.4	Choice of internal instrument	123
4.3.4.1	Choice of external test	123
4.3.5	Instruments chosen	123
4.4	English Ability Rating Questionnaire	124
4.4.1	The basic question	125
4.4.1.1	Cline of adequacy	125
4.4.1.2	Basic question and responses	126
4.4.1.3	Discussion of responses	128
4.4.1.4	Reliability of responses	129
4.4.2	The second question	129
4.4.2.1	Format of the questions	130
4.4.2.2	Interpretation of responses	131
4.4.3	Improvement in English	132
4.4.4	Trialling	132
4.4.4.1	Changes made	133
4.4.5	Further tuition in English	134
4.4.6	Type of course	135
4.4.7	Identifying information	135
4.4.8	Accompanying key	135
4.4.9	Questionnaire criteria satisfied	136
4.5	Communicative Proficiency Measure	137
4.5.1	Factors affecting the design	137
4.5.2	Content requirements	138

		Page
4.5.2.1	Oral communication sub-tests	139
4.5.2.2	Written communication sub-tests	140
4.5.2.3	Sub-test options	140
4.5.3	The interview	142
4.5.3.1	Introductory phase	143
4.5.3.2	Question phase	143
4.5.3.3	Monologue phase	145
4.5.3.4	Subject questioning phase	146
4.5.3.5	Assessment criteria	147
4.5.3.6	Analytic method	147
4.5.3.7	General criteria	148
4.5.4	The dictation	149
4.5.4.1	The text chosen	150
4.5.4.2	Principles of presentation	151
4.5.4.3	Length of 'chunks' and pauses	152
4.5.4.4	Presentation of the text	152
4.5.4.5	Criteria for assessment	153
4.5.5	The essay	155
4.5.5.1	The task-length	155
4.5.5.2	Time	156
4.5.5.3	Topic	156
4.5.5.4	Criteria for assessment	157
4.5.5.5	Analytic approach	157
4.5.5.6	Impressionistic approach	158
4.5.5.7	Use of both approaches	159
4.5.6	The reading test	159
4.5.6.1	Topics and length of text	160
4.5.6.2	Deletion rate	161

		Page
4.5.6.3	Passages chosen	161
4.5.6.4	Criteria for assessment	163
4.5.7	Trialling the measure	164
4.5.7.1	Trialling arrangements	164
4.5.7.2	Samples	166
4.5.7.3	Edinburgh sample	167
4.5.7.4	Aberdeen sample	168
4.5.8	Results	169
4.5.8.1	Dictation	169
4.5.8.2	The essay	169
4.5.8.3	The reading test	173
4.5.8.4	The interview	177
4.5.8.5	Overall results	182
4.5.9	Finalised Communicative Proficiency Measure	186
4.6	Conclusions	187
CHAPTER 5	THE FIRST INVESTIGATION	
5.0	Purpose	188
5.1	Collection of background data	188
5.1.1	Details of information collected	188
5.1.2	Identification of subjects	190
5.2	Distribution of the rating form	191
5.2.1	Preparation of the forms	191
5.2.2	Timetable	191
5.2.3	Effectiveness of the procedure	192

		Page
5.3	Characteristics of the sample	192
5.3.1	Age	193
5.3.2	Countries of origin	194
5.3.3	Sex distribution	195
5.3.4	Educational background	196
5.3.5	Studies in Britain	196
5.3.5.1	Subject areas	196
5.3.5.2	Levels of study	197
5.3.6	Pre-departure English	198
5.3.6.1	EPTB scores	198
5.3.6.2	BCSA results	201
5.3.7	Comparison of EPTB/BCSA results	202
5.3.8	Remedial tuition	203
5.3.8.1	Length of tuition	204
5.3.9	EPTB and BCSA sub-samples	205
5.3.9.1	Countries of origin	205
5.3.9.2	Age and level of education	206
5.3.9.3	Subjects studied in Britain	206
5.4	Analysis of the English ability ratings	207
5.4.1	Question 1	207
5.4.1.1	Sub-sample variations	209
5.4.2	Question 2	210
5.4.2.1	Use of the scale	211
5.4.2.2	Differences between skills	212
5.4.2.3	Notional total proficiency rating	214
5.4.2.4	Differences between sub-samples	216
5.4.3	Question 3	216

		Page
5.4.4	Question 4	217
5.4.5	Question 5	218
5.4.6	Question 6	218
5.4.7	Conclusions on tutor ratings	221
5.5	Conclusions on the first investigation	221
CHAPTER 6	DISCUSSION OF RESULTS OF THE FIRST INVESTIGATION	
6.0	Introduction	223
6.1	The Method	223
6.1.1	The product moment correlation	223
6.1.2	Contingency tables	224
6.2	Comparison by Correlation	225
6.2.1	EPTB sub-sample	225
6.2.1.1	EPTB, speaking and writing	226
6.2.2	BCSA sub-sample	227
6.2.3	Whole sample, speaking and writing	228
6.2.4	Discussion of the correlations	229
6.2.4.1	Adequacy of the obtained correlations	230
6.3	Comparison by Contingency Tables	234
6.3.1	EPTB sub-sample	235
6.3.1.1	Analysis of 2 x 2 contingency tables	236
6.3.1.2	Efficiency of the tables	238
6.3.1.3	3 x 3 contingency tables	240
6.3.1.4	Use of total skills ratings	241

		Page
6.3.2	BCSA sub-sample	242
6.3.2.1	Analysis of 2 x 2 tables	243
6.3.2.2	Efficiency of the tables	245
6.3.2.3	Analysis of 3 x 3 tables	245
6.3.2.4	Use of totals skills ratings	246
6.3.2.5	Conclusions	247
6.4	Other Variables	249
6.4.1	Remedial English	249
6.4.1.1	Correlations for the groups	250
6.4.1.2	Conclusions on the remedial groups	251
6.4.2	Country of origin	252
6.4.2.1	Correlations for the groups	253
6.4.2.2	Conclusions of the country groups	255
6.4.3	Subject of study	255
6.4.3.1	Correlations for the groups	256
6.4.3.2	Discussion of correlations	257
6.4.3.3	Education and TEFL	258
6.4.3.4	Other subject areas	259
6.4.4	Levels of study	260
6.4.4.1	Conclusion on levels of study	261
6.5	Conclusion	262
CHAPTER 7	THE SECOND INVESTIGATION	
7.0	The Purpose	264

		Page
7.1	The Method and the Instruments	264
7.1.1	The timetable	265
7.1.2	Collection of background data	265
7.1.3	English ability rating forms	266
7.1.3.1	Revisions to the rating form, questions 1 and 2	267
7.1.3.2	Revision of remaining questions	268
7.1.3.3	Accompanying letters	269
7.1.4	Communicative Proficiency Measure	269
7.1.4.1	Reading sub-test	270
7.1.4.2	Essay sub-test	271
7.1.4.3	Interview sub-test	273
7.1.4.4	Administration of the measure	276
7.2	The Sample	279
7.2.1	Characteristics of the sample	280
7.2.2	Age	280
7.2.3	Countries of origin	280
7.2.4	Education background	281
7.2.5	Studies in Britain	281
7.2.5.1	Subject areas	281
7.2.5.2	Level of study	282
7.3	Pre-departure English Assessments	282
7.3.1	EPTB results	282
7.3.2	BCSA results	283
7.3.3	EPTB and BCSA results compared	284
7.3.4	Remedial tuition	284
7.3.5	EPTB and BCSA sub-samples compared	285

		Page
7.4	English Ability Ratings	285
7.4.1	Responses to question 1	286
7.4.2	Responses to question 2	287
7.4.2.1	Total skills ratings	289
7.4.2.2	Sub-sample variations	290
7.4.3	Responses to question 3	291
7.4.3.1	Language tuition	292
7.4.3.2	Information on level of study	292
7.4.4	Further comments by tutors	292
7.4.4.1	Amplification of responses	293
7.4.4.2	Further comments on individuals	294
7.4.4.3	Comments on general problems	294
7.4.4.4	Comments on the questionnaire	295
7.4.4.5	Conclusions on tutors' comments	296
7.4.5	Reliability and validity	297
7.4.6	Conclusions	298
7.5	Communicative Proficiency Measure	299
7.5.1	The sample	299
7.5.1.1	Background of the sample	300
7.5.1.2	Pre-departure English	302
7.5.1.3	General conclusion on the sample	303
7.5.2	Administration of the measure	303
7.5.3	The Reading Test	304
7.5.3.1	Relationships betwen sub-tests	306
7.5.3.2	Conclusion on the reading tests	309
7.5.4	The Writing Test	309
7.5.4.1	Results	310

		Page
7.5.4.2	Assessment of the different levels	310
7.5.5	The Interview	314
7.5.5.1	Results	316
7.5.5.2	Discussion of the assessments	317
7.5.6	Results of different tests compared	320
7.5.6.1	Scores converted to categories	321
7.5.6.2	Overall results established	323
7.5.7	Conclusion	324
7.6	Conclusion	325
CHAPTER 8	DISCUSSION OF RESULTS OF THE SECOND INVESTIGATION	
8.0	Introduction	326
8.1	The Criterion Measures	326
8.1.1	Measures compared by correlation	327
8.1.2	Measures compared by attribution of categories	328
8.1.3	Measures compared by contingency tables	332
8.1.4	Summary of findings	333
8.2	Comparison of Pre-departure with Criterion Assessments	333
8.2.1	Comparison by correlation - tutors' ratings	333
8.2.1.1	Summary of findings	337
8.2.2	Comparison by contingency tables - tutors' ratings	337
8.2.3	Conclusion on the comparisons with tutors' ratings	341
8.2.4	Comparison using the CPM sample	341
8.2.4.1	Correlations between EPTB and CPM	342
8.2.4.2	Correlations between BCSA and CPM	344

		Page
8.2.4.3	Conclusions on correlations with CPM	345
8.2.4.4	Correlations between BCSA/EPTB and tutors' ratings - CPM sample	345
8.2.4.5	Comparison by contingency tables - EPTB and CPM	346
8.2.4.6	Comparison by contingency tables - BCSA and CPM	348
8.2.5	Summary of results of comparisons	349
8.3	Comparisons Affected by Different Variables	349
8.3.1	Female and Male groups	349
8.3.2	Remedial English groups	351
8.3.2.1	Conclusions on remedial English groups	353
8.3.3	Country groups	354
8.3.4	Groups according to subject of study	355
8.3.4.1	Relative levels of proficiency	357
8.3.5	Groups according to level	358
8.3.5.1	Conclusion on groups according to level	361
8.3.6	Summary of comparisons	361
8.4	Conclusions	362
CHAPTER 9	CONCLUSIONS	
9.0	The Context of the Study	364
9.1	The Current Situation	364
9.1.1	Characteristics of current types of test	365
9.1.2	Information on validity still required	366

		Page
9.2	Evaluation of the Results of the Investigations	366
9.2.1	General confirmation of the first two hypotheses	367
9.2.1.1	Workability of the procedures	368
9.2.2	General confirmation of the third and fourth hypotheses	369
9.2.2.1	Appropriateness of the criterion measures	371
9.2.3	Consideration of the fifth hypothesis	371
9.2.3.1	Trends across the two investigations	372
9.2.3.2	Conclusions on the fifth hypothesis	373
9.3	Discussion of Issues Arising from the Results	374
9.3.1	Establishing of tendencies	374
9.3.2	Notion of adequacy	376
9.3.3	Performance of ECSA	376
9.3.3.1	Consistency of EPTB performance	379
9.3.4	How important are proficiency tests?	379
9.3.5	Is there a case for ESP testing?	382
9.3.6	Is the testing of communication important?	382
9.3.6.1	The speed reading test	384
9.3.6.2	Communicative features to be tested	385
9.3.7	Is there an appropriate type of test?	385
9.4	Evaluation of the Approach	386
	References	389

VOLUME II

		Page
APPENDIX I	CRITERION MEASURES - PRELIMINARY VERSIONS	1
Section 1	English Ability Rating	1
	Trial version of English Ability Rating form	2
	Revised version of English Ability Rating form	3
	Key for completing English Ability Rating form	4
	Letter to tutors	5
Section 2	Communicative Proficiency Measure - trial version	6
	Dictation	7
	Essay	8
	Reading test passages	9
	Reading test passages - original texts	13
APPENDIX II	FIRST INVESTIGATION - SUPPLEMENTARY BACKGROUND DATA	16
Section 1	Supplementary Background Data - Whole Sample	16
	Table 5.4.1 Countries of origin by geographic area	16
Section 2	Supplementary Background Data by Sub-sample	18
	Tables 5.2.1 to 5.7.1	
APPENDIX III	CRITERION MEASURES - FINAL VERSION	21
Section 1	English Ability Rating	21
	English Ability Rating form	22
	Key to English Ability Rating form	23
	Letter to British Council Regional Directors	24
	Letter to university tutors and supervisors	25

		Page
Section 2	Communicative Proficiency Measure	26
	Reading test instructions and passages	27
	Writing test instructions and assessment scale	32
	Interview test content and assessment scale	34
	Letter to students	36
APPENDIX IV	SECOND INVESTIGATION: DATA AND RESULTS	37
Section 1	Background Data of the Whole Sample	37
	Personal data and test results sheet	37
	Tables 7.2 to 7.9	38
Section 2	Pre-departure Assessment Results - EPTB and BCSA sub-samples	43
	Tables 7.10 to 7.17	43
Section 3	Background Data of the EPTB and BCSA Sub-samples	47
	Tables 7.18 to 7.23	47
Setion 4	English Ability Ratings by Tutors. Results	51
	Question 1 - Tables 7.24 to 7.27	51
	Question 2 (skills ratings) - Tables 7.28 to 7.33	52
	Language improvement - Tables 7.34 to 7.35	54
	Language tuition - Table 7.36	55
	Correlations - Tables 7.37 to 7.38	55
Section 5	Comments Made by Tutors on the Ratings Form	56
Section 6	Communicative Proficiency Measure - Sample Background Data	78
	Pre-departure data - Tables 7.39 to 7.45	78

		Page
Section 7	Communicative Proficency Measure - Results	82
	Summary of results - Table 7.46	82
	Correlations,Reliability (cloze) - Tables 7.49 and 7.49.1	82
	Test results and correlations - Tables 7.50 to 7.54	84
Section 8	Selection of Essays by Level	88
Section 9	Content of Two Interviews	105
APPENDIX V	SECOND INVESTIGATION: FURTHER ANALYSIS OF RESULTS	107
Section 1	Tutors' Ratings and CPM	107
	Tables 8.1 to 8.3	107
Section 2	Pre-departure Assessments and Tutors' Ratings	109
	Correlations - Tables 8.4 to 8.6.1	109
	Cross-tabulations - Tables 8.7.1 to 8.7.6	112
Section 3	CPM Sample Data	118
	Tables 8.8. to 8.11	118
Section 4	Pre-departure Assessments and the Criterion Measures	120
	Correlations with CPM - Tables 8.12 and 8.13	120
	Correlations with tutors' ratings - Tables 8.14 and 8.15	122
	Contingency tables - Tables 8.16 to 8.19	124
Section 5	Pre-departure Measures and Tutors' Ratings, According to Different Variables	126
	By sex and remedial English - Tables 8.20 to 8.24.2	126
	By country of origin - Table 8.25	131
	By subject areas and level of study - Tables 8.26 to 8.27.1	133

CHAPTER 1 ENGLISH LANGUAGE PROFICIENCY TESTING

1.0 Language Examining and Testing - Background

The underlying theory and purposes of language examining and language testing are the same. No theoretical distinction between tests and examinations will be proposed. The distinction between the terms is largely a practical one for descriptive purposes. An examination is essentially a formal assessment which may contain a number of 'papers', individual tasks or tests. The use of the term 'examination' indicates emphasis on the person or persons conducting the assessment as opposed to the people or subject matter being assessed. It has very strong associations with subjective assessment, as in the phrase 'the oral examination'. The use of 'test' implies emphasis on the subject matter being tested, and is readily, though by no means exclusively, associated with objective forms of assessment, eg multiple choice tests. Tests are carried out in experiments and consequently the term has a more scientific connotation. 'Test' can also be used to refer to a battery of tests. The terms 'test' and 'testing' will predominate, therefore, and the terms 'examination' and 'examining' will be restricted to the contexts outlined above.

Foreign language testing is by tradition closely linked to foreign language teaching and public examinations. It is an activity which has been very much in the hands of language teachers. The preparation of short tests to assess students' progress during the school year as well as the setting and marking of end of term or end of year examinations is an accepted ingredient of the language teacher's work load. Tests are often additionally used as language exercises, particularly for reading and listening comprehension, and for developing a command of

the grammar. Moreover classroom teachers constitute the main body of examiners for the public examinations.

Great efforts have been made in the last twenty years to develop better language teaching techniques in the classroom, but there has not been a corresponding effort until very recently in the sphere of language testing. Testing practice in the schools has tended to reflect the testing procedures followed by the public examining bodies, while practice in adult education has tended to follow that of the schools in the first instance. However, in Britain the graded objective movement (Page, 1976 and Harding, Page and Rowall, 1980) has led to experimentation with more appropriate forms of foreign language testing techniques in the classroom, and the work of the Council of Europe (Trim, 1977; Van Ek, 1975) has focussed attention on possibilities for newer forms of testing for adults. The test prepared by Groot and Harrison (1978, unpublished) while not innovative in its construct could mark the beginning of a series of new language tests for adults in Europe.

1.1 Traditional Test Types

The following types of tests are traditional to foreign language testing.

(a) Translation (from L2 to L1 and from L1 to L2). Passages are normally connected prose and vary in length. At lower levels in L1 to L2 translation, single sentences are often given. L1 to L2 translation is often referred to as prose composition.

(b) Essay in L2. Usually on general subjects, but frequently on literary topics at the more advanced levels, eg final year of secondary school.

(c) Oral interview. Common ingredients are a reading passage aloud

in L2, and talking on everyday subjects. Marks awarded are frequently not incorporated into the general result of a language examination but reported separately.

(d) Dictation.

(e) Sentence completion or transformation - grammar.

(f) Questions on appreciation of the literature and culture.

Assessment in all these tests is primarily subjective.

Spolsky (1975: 3-4) characterises this stage in language testing as pre-scientific. In addition to enumerating the procedures (a) to (c) above he notes that marking was heavily subjective and that language teachers dominated language testing at that stage.

1.1.1 Vaughan James and Rouve (1973) were unable to find any clear statement of objectives relating to the examinations set for the General Certificate of Education, Advanced Level, in foreign languages in Britain. They attributed this to

> 'The refusal to make a sharp distinction between practical and cultural aims in modern language teaching'
>
> (op cit: 59)

In a survey of British-based examinations in English for overseas students Howell (1975: 40) was also unable to identify clear objectives or specific content in most of the syllabuses and concluded that it is only reasonable that it should be stated what such examinations aim to measure.

A further source of difficulty at this level of examining is that examinations such as the GCE Advanced Level in Britain, the Baccalaureate in France, and the College Boards in the USA have a dual function.

They act as measures to assess achievement at the end of secondary school and as entrance tests to higher education. Consequently they tend to be designed with the latter function in mind and to be geared to students intending to study a foreign language as a major or minor subject at College. Since it is only the minority of school leavers who intend to specialise in foreign language studies at university level, the need for a different end of secondary level test is more widespread and implies different kinds of examining procedures from those currently in use. A prospective employer or university tutor, as well as the individual concerned, surely needs to have clearer information on the ability of a student to use the foreign language in a variety of situations? A notion of language proficiency is required.

1.1.2 The characteristics of traditional language testing, the pre-scientific stage in Spolsky's terms, can be summarised as:

(a) lack of clear statements of objectives

(b) lack of a clear underlying concept of proficiency being tested

(c) almost completely subjective assessment

(d) control exercised by language teachers.

1.1.3 Language proficiency testing, and particularly testing proficiency in English, has developed against this general background. Whereas it can be argued that there has been no clear concept of language proficiency being tested, success in any of the public examinations cited above <u>implies</u> some degree of proficiency. This implied proficiency, and conversely the lack of

proficiency implied by failure, appears to have satisfied the British for a long time! However, a small number of teachers in some European countries earlier in this century were apparently dissatisfied with their English language examinations. The University of Cambridge Local Examinations Syndicate first set up the Certificate of Proficiency in English (CPE) in 1913, principally for foreign teachers of English. The examination included tests in phonetics, English literature and translation. The appeal of the examination was not very widespread at first, and there were only about 500 candidates per year by the late 1930s. But there was a demand for an examination at a lower level, and the Lower Certificate in English was introduced in 1939 (UCLES, 1973: 4).

1.2 Post-War Developments in Language Testing

The situation changed in the period after the second World War. Aid to nations recovering from the War, aid to the colonies both in their approach to independence and in the years following it, rapid scientific and technological advances, the increase in immigration to N America and Australasia together with more and more effective means of transportation and telecommunication are some of the factors which led to greater interchange across national and cultural boundaries and to the need for more people to speak more languages with a degree of proficiency that their formal education had not always prepared them for. In particular more and more people needed to use English.

The major impetus in developing language proficiency tests came from the USA through

(a) the applied linguistic research directed by Robert Lado at

the English Language Institute, University of Michigan, culminating in the completion of his doctoral dissertation in 1951* and the first examination for the Certificate of Proficiency in English in 1954 (University of Michigan)

(b) the work of the Harvard Language Aptitude project, directed by John B Carroll, a scholar with the rare distinction of being both a psychologist and linguist, which culminated in the Modern Language Aptitude Battery (Carroll and Sapon, 1959)

(c) the test developed by US government agencies, notably athe Department of Defense and the Department of State. Army Language Proficiency Tests were prepared for 31 different languages between 1948 and 1951. Modified forms have been produced since but their purpose continues to be a measure of the reading and listening proficiency of Defense Language Institute graduates (Petersen and Cartier, 1975: 119-130). During the period 1952 to 1956, the Foreign Service Institute developed a system of assessing speaking and reading using rating scales appropriate to the needs of government personnel engaged in international relations, principally diplomacy (Wilds, 1975: 29-38). Although details of these tests have not been made public until recently for security reasons, they are indicative of one area of considerable activity in language proficiency testing.

1.2.1 Spolsky refers to this second stage in language testing as the Psychometric-Structuralist trend since those active in the field drew heavily on the then current work in structural

---

* "Measurement in English as a Foreign Language with Special Reference to Spanish-Speaking Adults"

linguistics and psychometrics (op cit: 4-12). Language testing was seen as being a more precise and more scientific activity than before. There was a reaction by linguists and psychologists against the lack of precision in the content of traditional language tests and against the subjectivity of scoring. Detailed linguistic analysis was held to indicate where the problems lay and the language content to be tested. Lado stressed the importance of comparative linguistic analysis.

> 'The advance in English language testing came not from connected material but from concentrating on the language problems as they actually are. And we get closest to the language problems by a systematic comparison of the native language and the foreign language.'
>
> (Lado 1957: 4)

At the same time psychologists were stressing the need for validity and reliability in tests and were developing statistical techniques to achieve these qualities in tests. While Lado saw the value of these developments, he insisted nevertheless that linguistic and not statistical analysis should determine the content of a test, but that statistical techniques would contribute to the refinement of tests.

> 'Statistical treatment has its place in the refinement of the test, not in the selection of language problems.'
>
> (op cit: 5)

The notions of construct and content validity advocated by the psychometricians required that language be broken down into definable elements and factors that could be tested, as exemplified by Carroll's enumeration of four factors in language aptitude (Carroll, 1962). The work of the structural linguists and Lado's 'problems' approach coincided with these requirements, and language

came to be tested in terms of elements - phonology, grammar, lexicon - and integrated skills - listening comprehension, speaking, reading comprehension and writing. The theory of language as 'a system of habits of communication' (Lado 1961: 22) was also consistent with the current thinking of the behavioural psychologists. Objective scoring was also emphasised as being an important ingredient in the achievement of test reliability. Tests which measured individual 'problems' or items in the areas of phonology, grammar and lexis could readily incorporate objective testing techniques. This led to the increasing use of the discrete point item and the predominance of listening and reading tasks.

1.2.2 The developments in language testing during the post-war period 1945-1960 can be characterised as

(a) focus on discrete linguistic points

(b) bias towards testing the receptive skills and testing the linguistic elements through receptive skill tasks

(c) emphasis on greater test reliability and validity

(d) extensive use of objectively scored tests

(e) control exercised by linguists and psychometricians.

Since these developments arose because of the need for language tests in new situations, the traditional practices were still carried out on a wide scale. One effect of the new 'psychometric-structuralist trend' was to cause many individuals and public bodies to seek ways of improving reliability in the assessing of spoken and written production. Multiple marking, analytic marking, structured interviews and guided writing tasks were developed. These further developments were inevitable as the major new testing

insights did not extend significantly into testing the productive skills, nor to the testing of communication.

1.2.3 A typical English proficiency test constructed during the period under discussion consisted of

(a) listening comprehension

stimulus (spoken):	sentence, conversation, or monologue (short talk)
response:	selection of appropriate comment, continuation or reply from a number of written alternatives

(b) vocabulary

stimulus (written):	single word in context of a sentence or single word
response:	selection from a number of alternatives of either the most appropriate word to complete the sentence, or the most appropriate meaning conveyed by the word in the context, or a synonym. Many tests include both completion and recognition item types.

(c) reading comprehension

stimulus (written):	paragraph(s) of connected prose followed by questions. Up to five separate paragraphs on separate topics may appear in one test.
response:	selection from a number of alternatives of an appropriate response.

(d) grammar

stimulus (written):	one or two sentences, often in context of a conversation.
response:	selection or recognition of appropriate word or phrase from a number of alternatives.

The scoring of each subject would be objective. The total number of items would be in the region of 150 to 200, and the whole test would take from 1 to 2½ hours.

1.2.4 Table 1.1 provides an analysis of six tests developed in the 1950s and 1960s, together with the new form (1975) of the Cambridge Certificate of Proficiency in English examination. The tests, with dates given for the first form, are:

Mich: Test of English Language Proficiency, English Language Institute, University of Michigan, 1961

ALI/GU: Tests of English as a Second Language, David Harris and Leslie Palmer, American Language Institute, Georgetown University, 1961.

TOEFL: Test of English as a Foreign Language, Educational Testing Service, Princeton, 1964.

EPTB: English Proficiency Test Battery Short Form, Alan Davies, University of Birmingham, 1964, for restricted use by the British Council, London.

ELBA: English Language Battery, Elisabeth Ingram, University of Edinburgh, 1964, unpublished.

CGM: Test CGM 62, Mialaret et Malandain, CREDIF Paris, 1962, published by Didier.

CPE: Certificate of Proficiency in English, University of Cambridge, Local Examinations Syndicate, 1975.

The three American test batteries all conform to the test format outlined in para 1.2.3 above. The University of Michigan also requires candidates to write an essay if the certificate of proficiency is to be awarded. ELBA is the only British test which conforms to the format. EPTB does not contain a vocabulary test on the grounds of sampling difficulty. Monologues with questions did form part of the listening comprehension section of the full battery but were not included in the short form because of their low reliability (Davies 1965: 163). Phonetic discrimination or recognition at the segmental level is included in ELBA, EPTB and CGM, and at the suprasegmental level in ELBA and EPTB. These are characteristics of Lado's tests which have not been incorporated in the three American tests in Table 1.1.

Table 1.1 Analysis of six language proficiency tests

	CPE	EPTB	ELBA	ALI/ GU	TOEFL	MICH	CGM 62
Objective tests							
Listening							
Phonemes		√	√				√
Stress & Intonation		√	√√				
Sentence comprehension (+ reading m/ch)			√	√	√	√	√
Conversation comprehension (+ reading m/ch)		√^			√		
Monologue(s) comprehension (+ reading m/ch)	√				√		
Vocabulary							
Vocabulary single word definition			√		√		
Vocabulary in single sentence context/ recognition	√		√	√	√	√	
Style							
Choice of phrase or sentence				√	√	√	√
Acceptability					√		
Grammar							
Recognition m/ch	√	√	√	√	√	√	
Completion of (mod* cloze) passage	√	√					
Sentence completion	√						
Sentence transformation	√						
Reading							
Passage + m/ch	√√						
Para + m/ch			√√√	√	√	√	
Mod cloze passage *	√	√					
Intrusion in passage (speed)		√					
Subjective tests							
Reading							
Open question	√						√
Writing							
Directed writing	/						√
Essay	//					√	
Comment on text	/						
Dictation							√

	CPE	EPTB	ELBA	ALI/ GU	TOEFL	MICH	CGM 62
Oral							
Reading aloud	√						
Question on topics (pictures)	√						√
Free range	√						
Role play	√						
Proportion of tests objectively scored	40%	100%	100%	100%	100%	100%	60 ~~100~~%

* tests the same objective

√^ two objectives tested in one test

1.2.5 The Test CGM 62 did not owe its format exclusively to the developments in the USA but also to concurrent work in language teaching methodology at the Ecole Normale Superieure de Saint Cloud as exemplified in the course 'Voix et Images de France' (CREDIF, 1967). The authors of CGM pointed out that their aim was to provide 'indications', or information, on each individual by providing a language proficiency profile (Mialaret et Milandin, 1962: 8). Greater examiner/marker reliability was also aimed at through the provision of detailed instructions on how to administer the subtests and on how to assess them. In the case of oral and written production the tasks were closely controlled eg through picture compositions, and the examiner was told precisely what points to look for and score.

It illustrates well the trend away from traditional techniques to more discrete, objective and reliable items and subtests. It is the only test of the six tests of that period in the table that made extensive use of an examiner. It was the only one with dictation and the only battery in which the listening comprehension

was administered to the candidates individually. Indeed the whole battery could well be administered to one student at a time. A heavy burden was placed on the tester, but the special circumstances of CREDIF made this possible. The test was for use mainly at CREDIF in Paris and not overseas. Overseas students arriving in France who had not had their previous education through the medium of French were sent to the Centre who arranged testing sessions according to the demand. Examiners were mostly members of the permanent staff.

Another feature of the test was that scores were not reported as raw, percentage or standardised scores, as in the other tests, but as belonging to one of five levels. These levels were based on standard norm referenced scores for each test with a mean of 3 and SD of 1. The obtained levels of each candidate were then plotted on a diagram so that the recipient of the scores could see where the strengths and weaknesses of the candidate lay. The oral production subtest was difficult to conduct since the examiner had to follow the set format and assess the candidate at the same time, but the amount of interaction between tester and testee contributed to the construct and face validity of the test. The numbers of items in each subtest were low, averaging no more than 15, oral interview excepted, thus putting a heavy burden on the language sampling. The test did not contain a grammar subtest and was 60% objective.

1.2.6 Whereas the CGM test gave evidence of evolution from the traditional test forms, the CPE gave evidence only of adding newer methods and techniques to the old. It is nevertheless a comprehensive examination. Multiple choice listening comprehension,

grammar and reading comprehension questions were introduced. The interview and some of the writing was more carefully structured providing greater consistency in the types of expression tasks each candidate had to perform. The interview was incorporated as part of the examination. The traditional essay and questions on a passage remained. Translation, literature, and papers which before 1975 were optional were retained as optional additional papers not having any bearing on the final result. 40% of the new examination is objectively scored, but the examination takes 8 hours to administer in five separate sessions. The only results reported are three passing grades and two failing grades. No other information is given.

There remains a question mark over the efficiency of CPE. May there not be overkill? Is the proficiency being tested related to any particular situations? And how lacking in proficiency are those candidates that fail? These questions remain unanswered. Although the Certificate is recognised by the University of Cambridge and most other British universities for the purpose of exempting students from any further English requirement for matriculation, it is not widely used by candidates for that purpose. It is the only test or examination in Table 1.1 which is not administered primarily to assess a candidate's proficiency in English (or French in the case of CGM 62) for the purpose of studying at tertiary level through the medium of English (or French).

1.3 Current Developments in Language Testing

There is a further phase of development where tests of English for

academic purposes occupy a much less prominent position. It relates to the testing of functional language proficiency, communicative competence, and overall proficiency, stimulated by the inadequacy of proficiency tests based on structuralist grammar and discrete point items.

Claims have been made that paper and pencil tests can yield reliable information on a student's speaking or writing ability (Lado, 1961), but they have not been substantiated by the evidence. Although there is generally some positive relationship, there have been no clear indications that there are valid close relationships beyond a general language factor. New tests of communicative competence are now required for specific occupational and social needs, and integrative tests of a more general or overall proficiency are also being tried out. This third phase of development is referred to by Spolsky (op cit: 12-14) as 'the psycholinguistic and sociolinguistic trend'.

1.3.1 This current phase in language testing can be characterised as concerning itself with

(a) the notion of overall proficiency

(b) the functional use of language

(c) communication

(d) reliability and validity (a continuing concern).

It is too early to be more specific and it is not possible to point to any coordinated movement in proficiency testing nor to a typical test format. Research is still being undertaken. In the realm of theory it is the notion of overall proficiency that is receiving most attention as exemplified by the work of Oller (1979), Oller and Perkins (1980), Palmer and Bachman (1980), while in practice

money is being spent by national and international bodies for the development of measures of functional language ability as exemplified by the tests developed for employees of individual industrial companies (ELTDU, 1975), by the new test designed for students coming to study in Britain (ELTS, 1980), the new tests of English for communicative use designed primarily for young adults living in Britain (RSA, 1980), and the test designed for peace corps training programmes in the USA (Clark, 1980).

1.3.2 The basis of the notion of overall proficiency is the presence of a general factor of English language proficiency which, it is suggested, is indicated by the relatively high intercorrelations obtained between subtests and between batteries developed in the psychometric-structuralist phase - see Table 1.1 above. An examination of correlations between subtests of a given language proficiency battery will generally reveal intercorrelations of between .45 and .8. Intercorrelations reported for TOEFL range from .56 to .78 (TOEFL, 1973: 15); for ALI/GU from .46 to .78 (Harris, 1967: 14); for EPTB from .32 to .69 (Davies, 1965: 277); for ELBA* from .46 to .78 (Ingram, 1970). Inter battery correlations of .79 and .84 between TOEFL and ALI/GU have been obtained at Georgetown University (Harris, 1967: 15 and 1972: 4). correlations ranging from .71 to .90 between the Michigan tests of proficiency and EPTB in Singapore (Ibe, 1974: 15); and of .68 twice and .74 between EPTB and ELBA in Edinburgh (Pilliner, 1965; Moller, 1975).

---

* excluding intercorrelations with the stress recognition subtest in ELBA, as they were all well outside the range and lower.

A good performance on one or more of these batteries, however, does not necessarily mean that a candidate can undertake certain tasks satisfactorily through the medium of English, though the assumption that he can do so is reasonable. Nor does poor performance necessarily mean lack of ability to do so. What does success in one of the batteries in Table 1.1 indicate? The assumption that the sum of all the linguistic tasks demanded in any of the batteries adds up to either overall proficiency or ability to function in a variety of situations is questionable. It is after all an assumption.

1.3.2.1 This being so, alternative methods of measuring overall and functional language ability, with reliability, need to be found. One method being researched is based on two assumptions

> '(1) that there is such a factor as overall proficiency in a second language, and (2) that it may be measured by testing a subject's ability to send and receive messages under varying conditions of distortion of the conducting medium.'
>
> (Spolsky, 1967: 39)

Spolsky began investigating these assumptions and devised a test consisting of fifty sentences which were controlled for vocabulary and sentence structure, recorded on tape and to which white noise was added with varying signal to noise ratios. The subjects were to write down what they heard. Assuming that the sentences represented an acceptable sampling of the language - they were taken from other language tests - there were, and still are, certain practical difficulties. Firstly, it is not clear whether errors are attributable only to the noise level (Spolsky et al, 1968). Complexity of syntax seems to be the major difficulty according to another study (Gradman and Gaes, 1975: 5). Secondly, since all

the sentences are independent of each other, the context of each has to be decided by the listener. Thus an initial content word or proper name can be difficult to recognise, and the whole sentence may be imperfectly understood. Thirdly, the scoring is not simple.

In Indiana University, where the test was originally developed, subjects are presented with alternative sentences on the answer sheet and have to select the one they think they have just heard. A further refinement has been developed at Bar Ilan University in Israel. Subjects listen to a narrative with white noise added and at intervals the narrative stops. The subjects then read four sentences and select the one which they think they have just heard (Whiteson, 1972 and Seliger, 1975).

This type of test involves reading as well as listening under varying conditions, and is reliable, as well as easy to administer and score. Spolsky reported a correlation of around .6 between the original test and other language tests, and Whiteson reported only one unsatisfactory score for a group of 95 students who were divided into an advanced group and a less advanced group on the basis of scores obtained on the noise test. In a follow-up study in 1974 Gradman and Gaes (op cit) are satisfied that the test does determine overall proficiency in a group. In particular it has shown that a group of non-native speakers deemed to be of low intermediate ability in English do not perform particularly well on the test in comparison with native speakers. Yet in spite of this poor showing on the test they note (op cit: 10)

> 'Their academic performance was successful to the point that they were able to maintain their academic status, thus

> indicative that their functional knowledge of English was at least satisfactory for university level coursework.'

This finding is of interest even though based on a sample of only 29 subjects. It nevertheless complements findings elsewhere that low performance on an English proficiency test does not necessarily mean that a subject will be unable to participate satisfactorily in an academic course.

1.3.2.2 Dictation is another technique that has been re-examined recently. Used as part of the UCLA English as a Second Language Placement Examination (ELSPE) it has been shown to correlate very highly with the total score both when the part score is included in the total ($r = .94$) and when it is excluded from the total score ($r = .85$). Both correlations are higher than those of any of the other parts with total score (Oller, 1971 and Oller and Streiff 1975). In addition it has proved to be a powerful discriminator. A subject will perform well if his expectancies of what he is about to hear are confirmed -and those expectancies are presumably closely related to his language competence - and if the redundancy of the passage is not reduced. Since extraneous noise is purposely avoided, difficulties in perception will arise as a result of the speed of delivery and the length and syntactic complexity of the text being dictated. A proficient hearer/writer will retain the significant lexical and syntactic elements of what he has heard and supply the redundant features of the language. Scoring is not always straightforward. One phonological error may span two or more words eg 'lawn care' for 'long hair', but need not be complex. Conditions for administration have to be good and free from external noise. A

hall in the centre of a busy town on a hot afternoon is not a suitable location, for example.

Insufficient work has been done, however, to determine the validity of dictation. One experiment using carefully controlled dictation has yielded promising results (Fountain, 1974). Fountain constructed dictation passages consisting of five paragraphs. The key items were lexical items. The first paragraph contained ten key words at the 500 word frequency level. This paragraph was not scored. Each succeeding paragraph contained twenty key words of decreasing frequency levels. The average length of each dictated portion increased with each paragraph as did the syntactic complexity and the reading speed. Only the 80 key words in paragraphs 2 to 5 were scored. Morphological errors and single letter spelling errors were not counted. Phonological errors were counted. With a mean of 50 and SD of 18 the dictations were clearly discriminating well. The longer the dictated portions became the lower the facility of the individual item. The tests were taped and administered to non-native speakers following the three month pre-session English course at the English Language Institute, University of Wellington, New Zealand. The correlation with the pre-course test was .71 and with the post course test was .86. These results indicate the desirability of experimenting further with this format.

1.3.2.3 Another technique that is being studied as a measure of overall language ability is the cloze technique. It was originally developed as a measure of readability (Taylor, 1953) but has more

recently been used in experiments to measure reading ability in a second or foreign language (Anderson, 1971; Darnell, 1970) and in experiments to measure general language proficiency (Oller et al, 1972 and Oller, 1973). Both cloze tests and white noise tests make use of the feature of redundancy in language and both techniques distort the text by means of systematic deletion of words in the written text or by imposition of white noise over words in the spoken text. In this way redundancy is reduced and the subject's task is to write the text in its original form. Both tests can be considered as indirect measures of language proficiency. We are constantly conversing or listening to language against a background of noise, but not always white noise. It is claimed that the white noise test separates native speakers from non-native speakers, correlates well with other proficiency measures and discriminates well (Gradman and Spolsky, 1975). It should be noted, however, that the subjects reported on were undergoing English improvement courses and therefore may not have included high level non-native speakers of English.

Similar claims can be made for cloze procedure. In a recent study of current and alternative formats for TOEFL (Pike, 1973) the standard deviations observed for the cloze test indicated variability of scores matched only by the TOEFL listening comprehension section. Observed correlations between cloze and other subjective and objective tests were consistently between .6 and .8. The findings report reported in Chapter 7 below also show the discriminatory power of cloze, although the correlations with other measures were not so high - between .5 and .6. It has also been noted that highly proficient non-native speakers of English obtain cloze scores obtained by native speakers.

Cloze procedure also brings with it the problem of scoring. The most straightforward method is the exact word (ie as in the original text). The clozentropy method (Darnell, op cit) is less convenient, necessitating pre-testing on a large sample of native speakers and assigning logarithmic values to at least two decimal places to each response to each item. Thus every single response a subject makes must be awarded a specific value. Alternatively, acceptable responses can be decided in advance and only those responses will be given credit, but it is usual to overlook some unanticipated acceptable responses. In the Pike study (op cit) the exact word method and the clozentropy method correlated at above .96. Similar high correlations have been reported elsewhere.

1.3.3 The general proficiency tests outlined so far are considered inadequate on two counts. Firstly they do not directly measure the productive skills, and inferences drawn from performance on tests of the receptive skills or on test batteries can only be very general and not at all reliable. A good score on a listening test may or may not indicate that a subject can express himself adequately in spoken English. On the basis of the assumption that each type of language test is in fact also measuring a general language proficiency factor, it would be reasonable to expect an adequate performance in spoken or written English, particularly from those with a percentile rank of 85 or above. Similarly we might expect that those below the 15th percentile would not be able to speak or write any English with any degree of adequacy. But we can have no meaningful expectations of those ranked in between! Secondly, the tests do not sample language in appropriate contexts,

nor do they require subjects to perform tasks considered to be relevant in the light of their known future use of language. Thus while the test batteries considered above may well be appropriate for students who will be following more formal degree courses, they may be less appropriate for students planning to undertake agricultural research, for example, or research in other applied sciences. They may be even less appropriate for subjects who will need English in order to represent their company commercially or to undertake a study attachment in an industrial firm or in a professional or government office.

1.3.4 The Foreign Service Institute in Washington began to tackle this problem in 1952 in respect of the needs of State Department officials for languages other than English. By 1956 they had refined an oral interview technique which led to a candidate being rated according to proficiency rating scales for Speaking and Reading. Two interviewers conducted the test with a native speaker of the target language doing most of the questioning, and a native English speaker fluent in the target language observing and assessing. Candidates were required to play different roles or give instances of language used for particular purposes, eg giving instructions to a joiner about a piece of furniture that he is to make. Admittedly the situation was artificial but it was the closest to a real life situation that the testers came. The FSI procedure did not become widespread partly because of the inevitable security and secrecy that apparently has to envelop the activities of the State Department, partly because candidates were not representative of the foreign language learning population, and

partly because the prevailing interests at the time were focused on reliability and discrete point testing of language problems. It was also a costly method of testing, since it involved using full-time examiners so as to maintain as much consistency as possible. The procedure of the FSI tests has not changed fundamentally and is still in active use today.

The definitions of the FSI ratings are stated in general terms despite efforts to be as specific as possible, and they assume that raters have had considerable experience in interpreting criteria such as 'effective participation' in discussions, discussing 'with reasonable ease', 'broad enough' vocabulary. Examples of rating definitions for level R3 (Reading) and S3 (Speaking) are

> Minimum Professional Proficiency
>
> S-3 Able to speak the language with sufficient structural accurary and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics. Can discuss particular interests and special fields of competence with reasonable ease; comprehension is quite complete for a normal rate of speech; vocabulary is broad enough that he rarely has to grope for a word; accent may be obviously foreign; control of grammar good; errors never interfere with understanding and rarely disturb the native speaker.
>
> R-3 Able to read standard newspaper items addressed to the general reader, routine correspondence, reports and technical material in his special field. Can grasp the essentials of articles of the above types without using a dictionary; for accurate understanding moderately frequent use of a dictionary is required. Has occasional difficulty with unusually complex structures and low-frequency items.
>
> (Wilds, 1975: 37)

These definitions serve as one of the models from which new sets of criteria have been developed by different organisations and researchers eg British Council subjective assessment ratings, ALI/GU

oral interview ratings. All suffer from the weakness that they cannot possibly be sufficiently specific for any assessor. They act as guidelines for individuals or teams of assessors who then develop their own techniques to arrive at a particular rating. Thus once again reliability is dependent on the raters themselves.

1.3.5 To overcome the lack of specificity the language functions in question need to be defined, extensive samples of the language used in carrying out the functions need to be taken and criteria at different levels of ability established. Munby (1978) has worked out a comprehensive model for analysing language needs and making specifications. However, it is of only limited use in text construction since the language sample has to be determined. Munby's model does not provide criteria for sampling nor does it provide a model for generating discourse. Adequate sampling is of central importance to a valid language test while a valid test of language production must be able to take account of an individual's capacity to generate language appropriate to the situation. To overcome this difficulty a number of criteria, or objectives, eg ability to take a routine telephone message accurately, may be decided but once more the problem of quantity or variety of objectives to be sampled has to be solved. In addition there is a problem of testing techniques, which could become both lengthy and costly if the logical solutions of role play or setting up simulated situations are adopted (Jakobovits, 1970).

1.3.6 Large organisations in particular are facing up to these very real problems. A project to develop instruments to measure the

functional English language ability of employees at the World Bank in Washington has been reported (Krowicz and Garcia-Zamor, 1975). A framework of five levels is proposed. The different jobs undertaken by employees are listed and each is related to appropriate levels. The first three levels are divided into adequate, restricted and inadequate. Thus the minimum level required of a clerk typist is 1 - restricted - and the highest level necessary is 2 - adequate. By contrast a department director must be able to operate at a minimum of level 4, and ideally at level 5. Between thirteen and eighteen criteria are listed for each level, ranging from

> 'Can describe his/her present or more recent job in some detail'
>
> (Level 1)

through

> 'Can report another person's opinion on a familar subject'
>
> (Level 3)

to

> 'Can present different positions on an issue without committing him/herself personally to any of them'
>
> (Level 5)

Work in Canada to determine the functional ability of civil servants in their second language is based on the notion of generic skills (Smith, 1973). In Colchester language attainment scales have been established for a Swedish industrial company (ELTDU, 1975).

1.3.7 Functional language ability measures aim to test a subject's ability to communicate effectively in specific situations determined by the needs of a particular, restricted, social group. They do not usually aim consciously to measure overall language proficiency, but have certain features in common. Firstly, they relate functional language abilities to specific jobs or activities and

leave the consumer (often a departmental head) to decide the level of attainment required. Secondly, the population assessed is limited in number and to members of one social or employee group. Consequently the measures devised are very much 'in house' measures and unlikely to be applicable to subjects outside the group without modification. Thirdly, they are concerned with integrative language tasks and do not normally employ discrete point techniques. Fourthly, the reliability of the assessments is dependent upon personnel who are readily available and who have been trained in English testing. Fifthly, subjects do not usually obtain scores but are assigned a grade related to functional criteria.

1.4 Summary

1.4.1 The current state of English language proficiency testing is one of considerable activity on a broad front. The traditional methods of examining still obtain with essays, summaries and interviews. They are essentially examiner based. Criteria are not clear, assessment is mostly subjective and based on precedent and the experience and knowledge of the assessor.

1.4.2 This type of testing is now either supplanted or supplemented by test formats developed in the post war period which are language based. But language is seen as a series of sets of items or problems which can be tested discretely with objective scoring of the responses. Success is considered indicative of proficiency in the (foreign) language.

1.4.3 More recent work has been aimed at providing measures

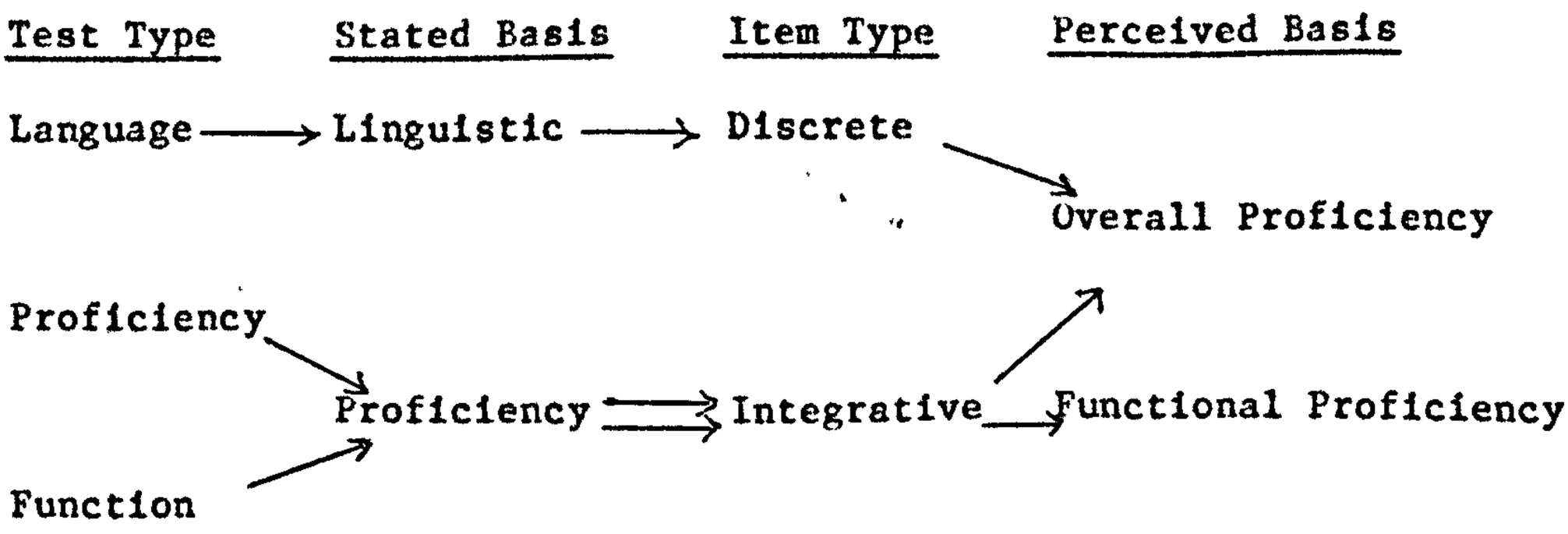
possessing greater construct validity. They can be distinguished as proficiency based. They are also 'language based' but the notion of overall proficiency is predominant. The basic testing techniques employed are not new, but the control over the linguistic content, and over the conditions under which the tests are administered and scored is new.

1.4.4 Overall proficiency is not what all test consumers seem to want, however. In practice many people want to know whether a set of subjects are able to perform specific tasks effectively in the medium of English. Measures are now being developed which are function based. Subjects are put in a particular situation and must make use of the language at their command at that moment to generate appropriate discourse. They may or may not possess overall proficiency. This will depend on the level and extent of the use of English required.

1.4.5 The distinction made between 'language based', 'proficiency based' and 'function based' tests may seem contradictory or artificial. This is deliberate. The theories on which they are based differ in their approach to language, and the methods of testing are different. It is reasonable to expect that they are measuring something different, even though they all claim to measure proficiency. In the case of the tests that have been labelled 'language based' and 'proficiency based', the differences in theoretical basis, item type and skill emphasis are clear, but the evidence suggests that both types of test seem to be measuring a general language ability, which in the case of English can be termed

'overall proficiency in English'. The type of test labelled 'function based' differs from the 'language based' test in theoretical basis, item type and skill emphasis. But whereas it shares some of the item types and skills emphasis with the 'proficiency based' type, it does not share the same theoretical basis. The evidence suggests that a strictly functional proficiency is being measured, but at the higher levels of functional proficiency it is probable that such tests measure overall proficiency as well. The comparisons are presented in figure 1.2.

Figure 1.2 Comparison of attributes measured by different types of test



1.4.6 Thus the 'language based' tests and 'proficiency based' tests, which claim to be measuring different types of proficiency, appear to be measuring the same factor of <u>overall language proficiency</u>. And the 'function based' test, which claims to start from a similar standpoint and to use similar testing techniques, appears to be measuring a different factor of <u>functional language proficiency</u>. The distinction, while looking like a distinction between receptive skill and production skill tests, is a deeper one.

The extent to which both the distinction and the differing types of test are valid will be discussed in the next chapter.

CHAPTER 2 THE PROBLEM OF VALILDITY

2.0 The Concept of Validity

The concept of test validity can most simply be expressed as establishing the extent to which a test measures what it sets out to measure. Only if it measures the ability or behaviour in question to the fullest possible extent can a test be said to be valid. Establishing a test's validity means

> 'inquiring whether the test measures what we want it to measure, all of what we want it to measure, and nothing but what we want it to measure.'
>
> (Thorndike and Hagen, 1969: 163)

Thus, if English language proficiency tests are to be classified as tests of overall proficiency or tests of functional proficiency, individual tests can only be said to be valid if evidence is obtained to support such statements as 'this test measures overall proficiency in English' and 'that test measures functional proficiency in English'. These assertions pre-suppose that such properties exist and that it is possible both to define and measure them.

In order to define these properties a theory of language is needed, as well as a description of English and a description of the purposes and situations for which the language will be used. The primary sources for these definitions and descriptions are linguistics and sociolinguistics. The discipline of psychometrics contributes the concepts of content, criterion and construct validity and underlines the need for an adequate theory and satisfactory criteria with which to relate the test results. Evidence for the validity of a proficiency test is dependent on the results of examining the test in the light of these concepts.

2.1 Types of Validity

In order to find evidence of validity Thorndike and Hagen identify three types of validity - content validity, criterion-related validity and construct validity (op cit: 164). In addition they acknowledge the notion of face validity - or what a test looks like - and suggest that it may be of importance only in making a test more acceptable or reasonable in the minds of the candidates (op cit: 166).

2.1.1 Cronbach (1960: 106) also identifies the same types of validity, but differentiates between two types of criterion-related validity, predictive and concurrent. Whereas Thorndike and Hagen devote their discussion of criterion-related validity almost exclusively to predictive validity, because of its importance in tests used to predict some future outcome or to select subjects for some future activity or employment, Cronbach, while agreeing the value of predictive validity in these situations, identifies a type of test for which concurrent validity might be more relevant. He suggests that where a test of a certain ability involves some inconvenient procedure, another test which is more convenient to administer may provide an acceptable estimate of the ability measured.

Horst (1966) also emphasises criterion-related validity, observing that the concept of validity is equivalent to the correlation obtaining between a psychological measure and a criterion (measure).

2.1.2 J B Carroll's work has provided an invaluable link between the disciplines of psychology, linguistics and language teaching. His work also serves as a reminder of basic principles for language

testers who lack a thorough background in psychology. In keeping with the thinking of the time he appreciated the need for content validity in English language proficiency testing, the predominant concern of Lado (1961), as well as the necessity for predictive validity.

> 'The validity of the test can be established not solely on the basis of whether it appears to involve a good sample of the English language but more on the basis of whether it predicts success in the learning tasks and social situations to which the examinees will be exposed.'
>
> (Carroll, 1961: 38)

For Lado there was no serious sampling problem since linguistic analysis could give insight into the problems of learning a foreign language and thereby highlight the elements to be tested.

> 'With the use of linguistic analysis and comparison of languages we are able to locate and describe the significant elements that will be most troublesome to a particular group of students. We are thus able to discuss content validity on more solid grounds than previously.'
>
> (Lado, 1961: 322-2)

The underlying theme is identification of problems of different language groups. These problems are likely to vary from one mother tongue group to another. Carroll, on the other hand, stresses the extent of a candidate's ability to function in a number of situations in the target language regardless of the native language. This, after all, is the desired future behaviour. Carroll is aware of the importance of, and difficulty in, defining the area of language from which the sample is to be taken and the consequent problems this poses for sampling (1961: 38-9).

2.1.3 In addition to highlighting validity in proficiency testing

Carroll pointed out the importance of test validity in his research into language aptitude and the prediction of success in foreign language training (Carroll, 1962: 98). More recently he has designated validity as one of the major 'persistent' problems of language testing, alongside those of scope (incorporating the sampling of linguistic tasks), efficiency and effect on teaching (Carroll, 1973: 6-17). He notes that validation against external criteria has not proved a sure guide to test validity and argues that more adequate analyses of the 'necessary competences, skills and discriminations' are required. This amounts to a call for greater emphasis to be placed on construct validity, which should lead in the first place to a reduction of evidence of invalidity in language proficiency tests. For, whatever the extent of research and development in foreign language testing 'it is in the nature of things that we will never be able to satisfy all our ideals and requirements in the field.' In comparing validation with accepting or rejecting a hypothesis he states,

> '...we can never accumulate sufficient evidence to prove that a test is valid beyond doubt, although we can often find evidence to indicate that a test is invalid and that it does <u>not</u> reflect what we want it to reflect.'
>
> (Carroll, 1973: 9)

2.2 Validity Studies of Language Proficienty Tests

Validity studies of language proficiency tests are not plentiful. Many important tests are still published without any validity studies being reported, as in the case of the Certificate of Proficiency in English (UCLES, 1973), while others like the Comprehensive English Language Test (Harris and Palmer, 1970) provide a considerable amount of normative

data in support of their claim to validity. Test constructors need to provide validity information before releasing their tests, since this information is required by the test consumers - candidates, employers, college admissions officers and scholarship awarding bodies - in order to help them to appraise the suitability of a particular test or battery.

Most recent publications on foreign language testing have devoted space to the discussion of validity, but in Lado (1961), Valette (1967 and 1977), Harris (1969), Clark (1972), Ingram (1974) and Heaton (1975) the techniques of item and test writing have predominated. The notion of validity receives a little more prominence and discussion in Davies' overview of the field (Davies ed, 1968 and Davies, 1978).

2.2.1 This lack of emphasis on validity is not altogether unexpected. Firstly, language teachers, for whom these books have mostly been written, are naturally concerned with the practical necesssity of producing short achievement tests geared in most cases to a specific syllabus or institutional situation. The validity of such tests will depend largely on the validity of the syllabus. The teacher's contribution is basically one of minimising sources of non-validity by producing tests with acceptable reliability that sample the syllabus adequately. Secondly, the practice of studying and reporting validity for English language proficiency tests coincides with the psychometric-structuralist trend in testing (see para 1.2.1 above). An English proficiency test - the Michigan English Language Proficiency Test - was reviewed for the first time in the Sixth Mental Measurement Yearbook (Buros ed, 1965). The seventh yearbook (Buros ed, 1972), contains reviews of TOEFL and

CELT. The appearance of these reviews indicates the higher degree of scrutiny to which language tests are to be submitted, and the reveiwers inevitably discuss the validity of the tests reviewed.

2.2.2 On a practical level, more and more test users are beginning to ask the questions which add up to 'How valid is this test?' The tests listed in Table 1.1 have all been in operation for some time, and their strengths and weaknesses have become apparent. Intuitive criteria applied by test users, such as face validity and the success - or lack of success - in communicating with native speakers, have led them to observe discrepancies between the English proficiency they expected of many non-native speakers as a result of performance on a test and the actual proficiency of those same speakers in specific situations. The current increase in students coming to English-speaking countries from countries where the tradition of English teaching is weak, where the culture is very different and where the immediate need is for increased technical education rather than academic study is a major factor in prompting this awareness. The demand for more relevant, more valid English proficiency tests is increasing.

2.3 Content Validity

Content validity, even though it has not in the past been acknowledged as such, has been the traditional concern of language testers both for public examinations and for the classroom. Traditional examinations have been concerned with the choice of texts, whether for translation, dictation, paraphrase or comment. Teachers have been concerned with the choice of texts and items relevant to the instruction that their

students have followed. Content validity is basic to Lado's work. Recent books on test writing assume content validity and offer test constructors a variety of techniques for item writing. Content validity is also at the heart of work currently being done in the field of teaching English for special purposes and related testing. It is at the heart of the functional proficiency tests such as the FSI interview and the test initiated by the General Medical Council in 1975 for administration to overseas doctors wishing to obtain temporary registration to practise in Britain. The test constructors and administrators of the latter were initially concerned that the language sampled should be relevant to the different clinical situations and should test proficiency in the different codes used according to different addressor/addressee relationships eg doctor/doctor, doctor/ patient.

2.3.1 Content validity, together with reliability, will ensure that a test adequately reflects the objectives and linguistic content laid down in a syllabus. In the case of a proficiency test, however, the test constructors themselves decide the 'syllabus' and the universe of discourse to be sampled. The sampling becomes less satisfactory because of the extent and indeterminate nature of that universe. Thus the evaluator looking for content validity is really assessing the test constructors' definition of proficiency. While it is conceivable and possible to have no opinion on test designers' specifications of English proficiency and to accept them as a basis on which to reach some conclusion about the validity of the sampling of their test, it is surely not a satisfactory procedure. The problem of sampling is linked to the problem of definition of the

domain of language proficiency. The more restricted the domain, the less problematic the sampling. But the wider the domain, the more unsatisfactory the sampling becomes, and the question arises 'What does it mean to know a language?'.

The literature on the complexities of the nature of language and of English in particular is now very extensive. While this presents a problem of selection of work that is relevant, more important is the fact that linguistics has not yet produced the proven theoretical bases on which to build definitions of what it is that we are testing. Lyons points to the inadequacy of current theories of semantics - 'No one has yet presented even the outlines of a satisfactory theory of semantics' (1968: 402) - and to the indeterminacy of grammar, which relates to the complexity of the rules necessary to account for all sentences generated, and to the problem of determining acceptability and unacceptability (1968: 152). This inconclusiveness does not make for the precision of definition of language proficiency that test validity imposes on the test constructor. In practice we are advised to accept for the time being the view expressed by Jakobovits (1970: 75).

> 'The question of what it is <u>to know a language</u> is not well understood and consequently the language proficiency tests now available and universally used are inadequate because they attempt to measure something that has not been well defined.'

Although too little is known about what knowing a language means, a lot is certainly known about English. This can present further problems in testing. What the test constructor decides a learner should know may well differ from what his teachers have decided he should know (Perren, 1970: 61). This leads back to the problem of sampling.

2.3.2 Uncertainty in knowing what to sample and how to sample is acknowledged by Peterson and Cartier (1975), and their solution, although a practical and temporary one is to relate proficiency closely to the content of the language courses in the Defense Language Institute which normally precede a candidate's attempt at a proficiency test. Thus the distinction between achievement and proficiency testing is blurred, since the objective of the syllabus is proficiency. This blurring of the achievement/ proficiency distinction seems undesirable. There would appear to be advantage in having dual sampling - one through the syllabus and one through the proficiency test - which would benefit the language learning and would act as a check on the syllabus itself, provided the syllabus did not encourage teaching towards the test (Davies, 1973: 18-26). It is also helpful to instructors and learners to be able to use proficiency tests that have wider application than to one special group. Cartier and Petersen's solution cannot take into account this wider application, since it is most unlikely that their sampling of eg English and of the linguistic tasks required of the candidate will coincide with that of many test users outside the Defense Language Institute. Their purposes are likely to be different.

2.3.3 Purpose is as crucial to proficiency testing as is defining language itself. Halliday has argued that 'language serves a wide range of human needs'; it is purposive in its function. Perhaps the only satisfactory response to the question 'What is language?' is 'Why do you want to know?' (Halliday, 1973: 9). Thus in English language proficiency testing we need to know why the subjects need

to demonstrate their proficiency in English. The more specific their purpose, the less problematic becomes the definition of the English to be sampled and how to sample it.

2.3.4 But in general terms content validity can be considered as a continuum. At one end the purpose for which the proficiency is required is clear and limited and the English to be sampled more readily defined. The probability of achieving content validity is at its highest. At the other end of the continuum the probability of content validity is at its lowest, as in the case of a test of overall proficiency with its general or experimental purpose, and where the domain of English to be sampled is imprecise. At this point the relevance of content validation for proficiency testing becomes less clear and the consideration of other types of validity is indicated.

2.4 Construct Validity

The low-yield end of the content validity continuum would seem to point to construct validity. Despite the general growing concern for validity in recent years the notion of construct validity has been almost completely neglected by language testers (Stevenson, 1975: 2). Davies (1965: 36) noted that 'language tests do not, as far as is shown, attempt to show construct validity'. Little has been done in the intervening period. Again he has noted the inadequacy of content validity in a proficiency test and the need to 'grapple with the terminal behaviour' through predictive validity or, better, construct validity (Davies 1965: 41, and 1967: 173). Carroll (1968: 47-54) attempted in greater detail than previously to establish a set of competences which go to make up

language proficiency and has more recently pointed out the limitations of validating against external criteria. This has led him to suggest that the examination of a test in terms of its construct validity is 'the only possibility of deciding on the validity of the test' (Carroll, 1973:10).

2.4.1 The consideration of what the construct of proficiency is brings us back again to the question ' what is proficiency in a given language? The same difficulties arise as when we attempt to define the domain of discourse from which our content is to be selected. While 'achievement' may be defined in relation to a syllabus or set of materials mastered in a formal learning situation, 'proficiency' is much more difficult to define. Briere offers as a definition

> 'the degree of competence or capability in a given language demonstrated by an individual at a given point in time independent of a specific textbook, chapter in the book, or pedagogical method.'
>
> (Briere, 1971a: 385)

Spolsky (1968: 94) concludes in an article on the problem of validation in language testing that our inability to be precise about the nature of language proficiency means that we are not yet in a position to question the validity of our items or tests. He recommends elsewhere in the article that

> 'we should aim not to test how much of a language someone knows, but test his ability to operate in a specified sociolinguistic situation with specified ease or effect.'

Clark (1975: 10) defines proficiency, in relation to proficiency testing, as

> 'ability to receive or transmit information in the test language for some pragmatically useful purpose within a real-life setting.'

Davies (1965) and Harris (1969) also characterise proficiency as the ability to do something.

The common ingredient of language proficiency tests would appear to be that a candidate demonstrate certain language competences, particularly self-expression and the ability to understand what is expressed by others. The notions of situation or setting, effect and purpose are also included. There is something imprecise about the term 'competence' as used by Briere, and Spolsky excludes the knowledge of a language, which is often characterised as linguistic competence. Because of the complexity of the nature of proficiency, discussion of it has not been extensive hitherto. But as Briere (op cit and 1971b) has pointed out, it is not clear whether the term 'proficiency' covers linguistic competence, communicative competence, or both. He argues for interdisciplinary work to come to an understanding of the nature of the variables contributing to communicative competence.

2.4.2 Such an analysis would build up a theory of language proficiency and put the task of construct validation within the realm of possibility. Cooper (1972) has presented an elaborated language model with its three dimensions of language skill, language knowledge and language variety. From this model it would be possible to test linguistic knowledge and make inferences as to linguistic competence. With the addition of contextual competence - knowledge of appropriate language use for the particular social context - it would be possible to test communicative knowledge and make inferences about communicative competence.

This kind of model has to be developed in terms of terminal behaviour as advocated by Davies (1965 and 1967), and in terms of competences, domains of discourse and contexts, as advocated by Carroll (1973). It may then be possible to establish effective construct validity criteria.

2.4.3 In the meantime most proficiency tests will be considered deficient in basic construct validity since they are essentially tests of linguistic competence with communicative ability limited to understanding only. In order to make good that deficiency the next step is the development of tests of production which can be scored with greater reliability than in the past. This is possible, though not easy. Jones (1975) has suggested specification of the language criteria to be evaluated, the training of judges, the construction of tests to be linguistically realistic and to sample the language rationally, proper administration, and accurate evaluation as the basic conditions to be met.

2.4.4 Few language proficiency tests have reported construct validity. Thirteen validity studies are reported for TOEFL, but only one relates to construct validity. This study was 'designed to determine the extent and areas of difference in English proficiency between American and foreign students', (TOEFL, 1970). 71 native American freshmen for whom scores on the American College Testing Program (ACT) were available took the TOEFL test. All had achieved scores at or below the 30th percentile on the ACT tests and were therefore representative of the lower range of ability of American college freshmen. Nevertheless the group's mean TOEFL

score was much higher than for all foreign students who had taken TOEFL between 1964 and 1969. The standard deviation was also much lower, as Table 2 shows.

Table 2.1 Difference in means and SD on TOEFL - 71 US freshmen and foreign students

	US group	All foreign 1964-69
TOEFL mean	622	487
SD	34	78

The evidence showed that the means (total and subtests) were high relative to those of the foreign students, that their range of scores was limited, that the score distributions were highly negatively skewed, and that many of the subtest scores were maximum or near maximum. This was particularly true of the listening comprehension, structure and vocabulary sections. The study noted that the writing ability section was the section which came closest to being discriminating for the American students. This was in line with expectations. Similarly the general results have shown that for the American students TOEFL is 'inappropriately easy', again confirming expectations.

While accepting the results of the study it is interesting to note that whereas the group of US students had all gained admission to a university, the performance reported for the foreign students relates only to those 'seeking admission to colleges in the United States'. Presumably a number of these students - and presumably mainly from among those in the bottom quartile - did not gain admission. If only the scores of those foreign students who had

gained admission had been considered, the mean would probably be higher and the standard deviation a little lower. Even so it is highly unlikely that the general conclusion of the study would have been affected.

The above study is not accepted by Chase (1972: 550-2) as evidence of construct validity as it does not point to a psychological construct in English proficiency. On the other hand it at least shows that the test does discriminate between native and non-native speakers of English, which is a necessary first step. EPTB (Davies, 1965) and CELT (Harris and Palmer, 1970) also report scores obtained by native speaker samples and thereby demonstrate the relative ease with which the native speakers can do the tests. They do not, however, claim construct validity on the basis of those results.

2.4.5 One further attempt at showing construct validity is the reporting of subtest intercorrelations. First these may show fairly low intercorrelations, in which case it may be claimed that each subtest is contributing something unique to the total score. This is argued in the case of TOEFL (TOEFL, 1973: 15) and intercorrelations are shown to be lower than the reliabilities of the part scores. Since those reliabilities range from .84 to .899, this should not be too difficult. Of the ten intercorrelations eight are in the range .64 to .78. The remaining two are .558 and .601, both involving the listening comprehension section. With the possible exception of that section, the extent of the uniqueness of contribution of the parts of TOEFL cannot be considered very important. The intercorrelations would seem to argue much more

in favour of a major common factor of language, or English, or proficiency.

2.4.6 Elsewhere intercorrelations have been used in attempts to show that one particular test or part or type of test is as efficient as others, or indeed as a whole battery. This is particularly true of the work done with the integrative tests. Spolsky et al (1968) reported that their 'noise test' compared favourably with the Indiana University test of English proficiency as a way of screening students with high and low proficiency. They claimed it would also result in time saving when administering it to the students. Darnell (1970) attempted to show that a test using the clozentropy technique yielded substantially the same information as TOEFL, and more efficiently. Oller (1971) and Oller and Streiff (1975) attempted to show the utility and validity of dictation as a measure of English proficiency by showing that dictation scores correlated more highly with total and part scores than any other part of the UCLA English as a Second Language Placement Examination.

He has extended this approach to determine the information yielded by both dictation and cloze tests when compared with that yielded by TOEFL scores obtained by a sample of Iranian students. The conclusion that 'it seems that the total score on TOEFL provides little interpretable information not contained in the cloze test, the dictations and the LC section of TOEFL' (Oller, Irvine and Parvin, 1974) is somewhat tenuous. Firstly the superiority of dictation does not emerge as it did in the UCLA study, and secondly the correlations being compared in the case of cloze are in the narrow

range of .67 and .81. These correlations are of TOEFL subtest scores and cloze and dictation test scores with total TOEFL, exclusive of the subtest which is being correlated. Thus the structure subtests correlate .77 with total TOEFL minus the structure subtest. Cloze correlates .78 with the same modified total. This difference of .01 shows no particular superiority of cloze over the structure subtest. The other differences quoted are of .14, .09, .02 and .02, which can hardly be significant. It would appear that cloze and dictation make a contribution equal to that of each of the other subtests.

2.4.7 The above studies are all based on small population samples. Much more evidence has to be gathered before the correlations become meaningful. On the other hand much work remains to be done in analysing and specifying the criteria being looked for in proficiency tests and the factors underlying that proficiency. Linguistic factors will not be sufficient. Sociolinguistic factors must be accounted for as well. They can be summarised as the 'social context' and relate to functions of language in specified contexts. The knowledge a speaker has of his native or other language cannot be characterised solely in terms of language usage but also in terms of language use and the contexts in which it is used. Since proficiency has always been typified as doing, functioning or operating, we may consider the construct of proficiency tests from the starting point acknowledged by Hymes (1972: xix).

> 'an adequate theory of the functioning of language would not "start" from either language or context, but would systematically relate the two within a single model'

It could be that a more refined theory would favour 'starting' either from language, as it does now - by default of context - or from context, as Hymes advocates. The answer is likely to lie in a basic construct in which language is interrelated with context, which can be modified to put more weight on either language or context according to the purpose of the test.

2.5 Criterion Validity

Criterion-related validity is a more easily demonstrable form of validity than content and construct validity and equally favoured by both test constructor and test consumer. If an independent criterion, for example success or failure in a degree course, exists or if there is another procedure measuring the same behaviour as that measured by the test, it is possible to compare the results of the two measures and determine the extent to which they agree or co-vary. This is normally determined statistically by correlation or by expectancy tables.

2.5.1

Reference has already been made to the distinction between concurrent and predictive criterion-related validity (2.1.1 above), and the language testing textbooks, beginning with Lado (1961), all make this distinction. The differences between the two are primarily of time and function but not necessarily of type.

The concurrent criterion is applied, as nearly as possible, at the same time as the measure being developed. But there will be a lapse of time before a criterion designed to establish whether a test has predicted satisfactorily can be applied. The period between the application of these two measures may be only a few weeks or as long as two or three years.

The functions of the criteria are also different, as their names imply. For concurrent criteria, measures are required that purport to be measuring the same abilities as the test being validated. The test and the criterion have to relate to the subjects' abilities at the same period of time in order to minimise the effects of any language learning achieved after taking the test. Proficiency in a language usually implies use of the language for some task or behaviour in the future. Consequently the predictive criterion has to relate to the terminal behaviour anticipated, or for which a subject is to be selected.

Predictive and concurrent criterion measures do not necessarily differ in type. One of the most frequently used criterion instruments is the teacher rating of subjects' ability. This is commonly used as a concurrent criterion, as in the development of EPTB and Chaplen's test (Chaplen, 1970). But such rating can be used as a predictive criterion measure also, as in Ibe's study in Singapore (Ibe, 1974). Another measure, teachers' grades, is frequently used as a predictive validity criterion, but it could be argued that there is little difference between teachers' grades and teachers' ratings. In practice, grades tend to be awarded by the teacher according to his own or his institution's criteria, whereas rating scores tend to be worked out by the researcher or other external body. An example of the use of a criterion measure for evidence of both concurrent and predictive validity is to be found later in the present investigation. The specially designed communicative proficiency measure was designed firstly to provide evidence of concurrent validity for the tutor's language ability

rating form and secondly to provide evidence for the predictive validity of the EPTB and subjective assessment procedures being investigated.

2.5.2 Concurrent validity may be examined in a variety of situations. If information on subjects is obtained as a result of cumulative assessments over a period of time, the validity of the final assessment can be confirmed, or questioned, as a result of administering an established proficiency test at the end of that period. An example of this procedure is reported by the investigator elsewhere (Moller, 1975). The English Proficiency Test Battery, Short Form, verion C, was administered to 48 overseas students at Atlantic College in South Wales who represented about half the number of first year students who were non-native speakers of English and who had been subject to a variety of assessment procedures during their first two months at the College. They had just embarked on a two year course leading to the International Baccalaureate. The sample was ranked in order of achievement based on the College assessments including performance on a test given on entry, essays and teachers' reports of progress. The rank order correlation between these assessments and performance on EPTB was .88, indicating a high measure of agreement based on two very different procedures.

The same study also provides a good example of the reverse process - the validation of a proficiency battery by a more comprehensive set of assessment procedures. Both parties to this study were satisfied with the outcome. Each party was looking for evidence

of validity for their particular measure, but each was prepared to consider the other party's measure a valid criterion. In the event their judgements seem to have been justified and evidence of concurrent validity for both measures was obtained.

2.5.2.1 Ingram (1974: 330) notes that the best criterion for a language test is teachers' ratings. This procedure is particularly appropriate for achievement tests and for proficiency tests where the population being tested is undergoing instructiuon and where it is accessible. It is not always possible to use this method for tests which draw on large populations, many or most of whom are not following instruction. Seven concurrent validity studies are reported for TOEFL (TOEFL, 1970) but only one used ratings as a criterion. These ratings or categories are partly based on performance on further tests and partly on judgements of ability to pursue academic courses. It was not clear to what extent TOEFL scores influenced these judgements - correlations ranged from .76 to .87.

Teachers' ratings were used as criteria for EPTB reported by Davies (1965 and 1967), for ELBA reported by Ingram (1970) and for CELT (Harris and Palmer, 1970). The individual groups of students (overseas students in UK) for whom ratings were obtained were mostly quite small. For EPTB the size of the groups varied from 10 to 55. Correlations significant at the 5% level, ranging from .46 to .84 were obtained for six of the groups (N = 160). Correlations for the remaining four groups were not significant. The overall correlation was .475 (N = 254). All groups with a non-significant

correlation were designated 'mixed groups' whereas three of the groups for whom significant correlations were obtained were groups of students following specialised courses in the general field of education. The four groups for whom significant correlations are reported for ELBA were also small and following specialised courses - 2 groups of students at a Norwegian Business school (N = 37), European technology students following a pre-university English course (N = 43) and a group of students of education (N = 19). Correlations from .61 to .91 were obtained.

Apart from individual groups of students the results of this method of correlation are not very encouraging. The rating system is bound to be quite crude, being based on minimal definitions of certain levels or, as with TOEFL, worked out according to the type of academic and English learning performance students are assigned to. CELT used a simple criterion of placement in English classes that had already been affected as a result of tests within the institution. When CELT was administered to the group, which had been divided into high, mid and low level groups a month previously, performance on CELT confirmed the grouping, the mean scores for the middle group being approximately half a standard deviation below the high group and half a standard deviation above the low group. Correlations between corresponding test sections, eg English structure sections in TOEFL and Michigan, ranged from .47 to .75, while correlations between test total scores ranged from .68 to .89. Inter-section correlations between CELT and TOEFL have also been reported (Harris and Palmer, 1970) and range from .71 to .83. Correlations of .79 and .84 are reported for the ALI/GU total scores

with TOEFL total scores. These correlations indicate that these proficiency tests are of comparable validity and are testing the same or similar language behaviour, but they do not indicate that this language behaviour is necessarily 'true' or 'absolute' proficiency.

2.5.2.2 The most commonly adopted criterion for important proficiency tests, particular in the United States, is another English language test. Thus TOEFL reports four studies using other proficiency tests as criteria, one study using 'cloze' and one study using writing on four different topics. The other proficiency measures used were the ALI/GU test, ALI/NY test, Michigan Test of English Proficiency and the UC Berkeley test (TOEFL, 1970).

When EPTB was developed there was no comparable proficiency test available in Britain. Correlations between EPTB and other proficiency tests have been reported more recently, however, .68 with ELBA (Ingram, 1970), .76 with ELBA (Moller, 1975), and from .71 to .90 with the Michigan tests of Aural Comprehension and English Language Proficiency (Ibe, 1974). These correlations follow the pattern established by the American tests quoted in the preceding paragraph, but fewer such studies have been carried out in Britain.

2.5.2.3 The major problem with criterion validation is the nature of the criterion itself.

> 'In most validation studies, the problem of criterion selection and development cannot be fully resolved. Even when ultimate criteria can be agreed upon and clearly stated, feasibility constraints typically require compromises of such nature that

> the criterion measures adopted are only relatively more direct than "those being validated".
>
> (Pike, 1973)

Thus using other proficiency tests is really a circular process since they are not more direct measures than the test being validated. However, experimental conditions may be conducive to constructing and using measures which may be more direct. Thus the Pitcher and Ra study (1967), in which an essay writing criterion test was developed and scored with a reader reliability of .92 for a sample of 310 students, showed that the TOEFL writing ability section - together with the structure section - correlated more highly with the writing criterion than the other sections of TOEFL did.

Pike (op cit), in his validation study of TOEFL formats, used more direct measures - interview and essay - and 'open-ended objective measures' - the Hunt rewriting task and cloze procedure - as criterion measures. In order to make the direct measures as reliable as possible, the interviews followed a set pattern, were taped, and rated independently by three listeners following a detailed rating procedure. The essays were also rated by eight independent markers according to four scales. Because of the size of the samples (98 Peruvians, 145 Chileans, and 199 Japanese) marking sessions were carried out on set days with 21 raters present, all of whom had followed a common orientation. As a result of these procedures reliabilities of .89 and above were obtained for all the samples for both interviews and essays with one exception - the interview for the Japanese sample in which reliability was 25 approximately .74. In the final analysis Pike

reported correlations (after correction for attenuation) of .75 and .84 between the interviews and listening comprehension (all groups), .86 to .88 between interviews and structure (Peru and Chile, lower for Japan), .77 to .93 between essays and writing ability (Peru and Chile, lower for Japan), .86 to .98 between essays and structure (Peru and Chile, lower for Japan). These results give some evidence for the validity of the listening comprehension and structure sections of TOEFL and also suggest that the writing ability and structure sections yield much the same information on a student's writing ability.

The method of criterion validation adopted by Pike, ie the use of specially constructed multiple choice objective, open-ended objective, and subjective measures, has the disadvantage of involving further tests, which he recognises. But it does to some extent overcome the problem of similarity of measures when using another proficiency test as criterion measure, and the problem of consistency when using ratings by a variety of teachers as a criterion measure. This method has the further merit of coming closer to construct validation.

2.5.3 The problem of criterion is just as persistent with predictive validation. The fact that a lapse of time is necessary between the measures being validated and the application of the criterion measure does not lessen the problem. If the time lapse is quite long, a subject's proficiency may have changed either as a result of increased learning and exposure to the language or because of lack of further contact with the language. A study of

overseas students going to the USA for study (N = 479) has shown correlations of between .6 and .68 between ALI/GU scores obtained overseas and scores obtained on arrival in US (AACRAO/AID, 1971).

2.5.3.1 The criterion measures used may be the same types of measure as used for concurrent validation but may also take the form of a measure of some other activity for which proficiency in the second language is essential. The most common situations for which predictive validation is required are those situations in which the subjects are to use English, or another language, as a medium for further study or for carrying out professional duties. In these cases measures of performance in their studies or in their professional occupation in the English-speaking situation may be taken as criterion measures. Such measures, however, will not be exclusively language measures. They will essentially be measures of a number of abilities of which language will be one. Satisfactory performance may be due more to skill and expertise in the subject being examined than to language proficiency and may be achieved in spite of poor language proficiency. Similarly, inadequate knowledge of a special field cannot be made good by a near native or native command of the language. Adequate English proficiency should enhance a student's chance of academic success and eliminate one source of weakness which could lead to failure.

2.5.3.2 The few studies in predictive validity that have been carried out have almost all used non-linguistic criteria. TOEFL reports five predictive validity studies (TOEFL, 1970). Four of these used the Grade Point Average (GPA) as the criterion. In US

universities grade points are awarded for each semester course at both undergraduate and graduate level. The grades and points awarded are according to the following scale - A = 4, B = 3, C = 2, D = 1, F = 0 - and are awarded on the basis of academic quality of the work submitted by the student. Points are totalled at the end of each semester and the average is computed. GPA has been criticised in the USA on the grounds that the range of grades is restricted and that the grades do not reflect any research or professional skills that may be important at the graduate level in particular.

The largest study reported (Maxwell, 1965) involved 238 students, and the correlation betwen TOEFL and GPA for the entire group was .17. This compared with a correlation of .11 between GPA and the University of California at Berkeley English tests for the same group. Subgroup TOEFL/GPA correlations varied from .02 to .58. The latter was obtained by the undergraduate subgroup (N = 47), and a correlation of .50 was obtained by the Middle Eastern subgroup (N = 27). The correlation for the graduate group (N = 192) was only .02. Further TOEFL/GPA correlations reported were

.26 N = 67 at University of Washington, 1966
.26 N = 61 " " "
.42 N = 88 at Massachusetts Institute of Technology (Chalmers, 1964)
.31 N = 50 at Fresno State College (Domino, 1966)

The fifth study reported was carried out in Hong Kong and reported by Byers (1969). It did not use academic success as the criterion but examined the relationship between grades obtained in the English examinations of the Hong Kong School Certificate and Certificate of Education. The study appears to be part concurrent validity

and part predictive - in reverse - since the object was the prediction of TOEFL scores. 2,026 students in the Anglo-Chinese (English medium) secondary schools formed one group and 296 students at Chinese-medium middle schools the other. Correlations of .67 and .70 respectively were reported, but this study serves to show the relationship between the Hong Kong exams and TOEFL and not TOEFL and subsequent performance in English. The above studies were all reported during the five years following the appearance of the first form of TOEFL and did not form part of the original validation procedures.

2.5.3.3 Predictive validity was recognised as being important by Davies (1965), and a predictive validity study was carried out as part of the development of the English Proficiency Test Battery. Examination results were obtained and used as the criterion. Rank order correlations between exam results and EPTB Short Form (using Fisher's z) varied from -.2 to +.7 with a total correlation of .455 (N = 208). However, none of the subgroups exceeded 31 in number and consequently all but one of the subgroup correlations under .5 failed to be significant at the 5% level. The correlation was almost identical to that of the concurrent criterion, which suggests that teachers may find it difficult to make much distinction between language proficiency and academic performance.

In her study of overseas students and nurses in Britain Sen (1970) administered EPTB and obtained final examination results for 1,010 students. The criterion was a simple pass/failure classification. Results were analysed for ethnic subgroups and again for

'qualification' or level of study subgroups. Discriminant equations comprising weighted average scores were established to predict the final course result on the basis of an individual's scores on the separate subtests of EPTB. Sen found that for practical purposes the test scores did not provide a useful guide to final academic performance.

> 'Although there were a few significant results, more than 35 per cent of students would be misclassified if the test were used to predict success or failure.'
>
> (Sen, op cit)

She further concluded that 'the extent of the use of and familiarity with the English language has little relevance to their final performance'. What Sen seems to overlook is that while the fact of very good English proficiency does not necessarily mean a good measure of academic success - or even any success at all - poor proficiency in English increases the likelihood of failure.

2.5.3.4 The purpose of ELBA is to distinguish students who will have varying but serious language difficulties in pursuing their higher education studies in the medium of English from those who will not. Consequently a follow-up or validity study has been in progress since 1968 at the University of Edinburgh (Ingram, 1973). Once again 'academic success' has been used as the criterion. But Ingram has defined this in two ways - firstly as success in the first available examinations, usually held at the end of the first term, and secondly as success in the examination at the end of the academic year. She has not reported correlations but has provided tables comparing ELBA score with the outcome of those examinations. A clear trend has been established. 'The higher the score on ELBA,

the greater the chance of passing academic examinations' (Ingram, op cit). ELBA has reported means of 70% and 66% with a reported SD of 14%. Of the students achieving scores of 80% and over, 88% passed their first examination and 89% the end of year examinations. For those students gaining scores of 70-79% on ELBA the success rates were 68% and 90%. Those students with higher scores would appear to be at an advantage at the very beginning of their studies, but the gap between them and those in the range 70-79% is eliminated by the end of one year. For students with scores between 50% and 69% the success rates were 50% and 66% respectively, while those students with scores of less than 50% had success rates of 30% and 57%. This latter group is clearly 'seriously at risk' and has little more than a 50% chance of final success. The overall pass rate at the final examinations was 79% (N = 392). In terms of prediction 81% of the students who gained scores of 50% (ie just over one SD below the mean) and above were successful in their end of year examinations.

This figure compares very closely with the findings of Davies (1965) who established the optimum predictive cut off for EPTB of 34.0, and SD below the mean score of 40. According to expectancy tables nearly 80% of subjects with EPTB scores of 34 and above gained a pass in their final examinations.

2.5.3.5 Evidence in accord with the findings of the previous paragraphs is to be found in a report on the selection and placement of recipients of training awards from the Agency for International Development, US Department of State (AACRAO/AID, 1971).

Significant correlations between English language tests and criteria of GPA and a specially computed Achievement Index (AI, based on GPA and credits obtained), are reported but are too low to be of much practical value in predicting an individual's success. Table 2.2 gives the correlations between scores on TOEFL and ALI/GU and end of first year criteria.

Table 2.2 Correlations between scores on ALI/GU and TOEFL and end of first year criteria (USAID sponsored students)

	Criteria			
	GPA (undergrad)	AI (undergrad)	GPA (grad)	AI (grad)
ALI/GU	.23	.32	.14	.30
TOEFL	.25	.36	.19	.33
	(N = 265-410)		(N = 390-510)	

On the other hand the overall performance of all the students was satisfactory (N = 953). Median first year GPA for graduates ranged from 3.0 to 3.2 and for undergraduates from 2.6 to 3.1. The report noted

> 1. There is a trend for the subgroups better prepared in English to achieve better grades in the first year. This trend is more pronounced for graduates than undergraduates.
>
> 2. It should be noted, however, that all subgroups are performing at a generally acceptable level, even where English background would seem to be deficient.
>
> (op cit)

2.5.3.6 The low predictive value of English proficiency tests is not reported universally, however. Two studies report more satisfactory predictive power. Ibe (1974) investigated the language proficiency of participants attending 4-month intensive courses in Teaching English as a Second or Foreign language at the Regional

English Language Centre, Singapore. Each course group consisted of an average of three teachers of English from at least eight different countries. The average course numbered 25 participants. Although the purpose of the study was to examine the changes in English proficiency in the groups by the end of the course, a predictive validity study was carried out with one of the course groups. The students were given EPTB, and the Michigan Aural Comprehension and Language Proficiency Tests at the beginning of the course. The scores obtained were correlated with numerical scores representing achievement in four of the sub-courses as well as with the overall course mark. Significant correlations were obtained in eleven of the twelve correlations between the proficiency tests and the sub-courses, ranging from .29 to .65. Correlations with overall mark were:

EPTB	.69
Michigan Aural Comprehension	.57
Michigan Language Proficiency	.80

These correlations are high, particularly with EPTB and Michigan Language Proficiency, which is a grammar, vocabulary and reading comprehension test. It should be noted that the subjects were practising teachers of English with mean scores above the standard means for these tests and were following a specialist course in TEFL/TESL.

2.5.3.7 Heaton and Pugh (1974) reported findings which they considered 'confirm the popular view that language ability is important for academic success'. Overseas students took an English proficiency test on arrival at Leeds University. The mean for the test was reported at 70% with an SD of 20, the sample numbering 354.

Final results were reported for their academic course as follows: Pass/Fail/No result. Two thirds of the sample obtained passes, 10% failed their course, and the remainder did not have a result for varying valid reasons. The first finding was that there was a significant difference between the mean English proficiency scores of the passing group and of the group who failed - at both undergraduate and graduate level. The second major finding was that there were highly significant correlations between the scores on the English proficiency test and examination performance for the undergraduate group and some postgraduate sub-groups. The correlations obtained (all significant at the 5% level unless indicated) were

Undergraduate group	(N = 65)	.537
Graduate group - all	(N = 180)	.164
Graduate Arts MA	(N = 21)	.65
Graduate Econ and Soc Diploma	(N = 13)	.059 not significant
Graduate Education Diploma X	(N = 16)	.595
Graduate Education Diploma Y	(N = 11)	.602
Graduate Education Diploma TEFL	(N = 38)	.667

It is to be noted that the highest correlation was obtained for the students following a diploma course in TEFL. These students were similar in background to those in Singapore, referred to in 2.5.3.6 above, and the correlation between language proficiency scores and final results is of the same order as those reported by Ibe (op cit). The high correlations for students following TEFL courses were not entirely unexpected, however. It may be that those students with more advanced proficiency in English have greater confidence in their ability to handle the methodology and applied linguistics courses which make up the programmes and consequently perform better in the final evaluation. The higher correlation for the undergraduate group and the low correlations for the

graduate group as a whole were consistent with the findings of Maxwell (1965), reported in para 2.5.3.2 above.

2.5.3.8 Gue and Holdaway (1973) gave further evidence of the limited validity of language proficiency tests. Four groups of Thai teachers were sent to the University of Alberta to follow a one year course leading to the Graduate Diploma in Education during the four academic years from 1967 to 1971. The Groups consisted of 37, 37, 24, 25 students respectively. They were selected in Bangkok in 1967 and 1968 by local test procedures which included panel ratings. TOEFL (summer) was administered to the students one week after arrival in Canada. They then followed a two month intensive English course, and TOEFL (autumn) was again administered about a month after the end of the course. GPA scores - in the range 1 to 9 - were obtained for the students at the end of their course (5 being generally considered to be the pass grade). The mean summer TOEFL total for all groups was 424 (3/4 SD below the norm mean). The mean GPA was 6.6 Correlations of .49 and .59 were reported between the summer TOEFL and GPA and the autumn TOEFL total and GPA respectively. Mean correlation coefficients for the TOEFL subtests ranged between .34 and .55. Stepwise multiple regression analyses were performed using GPA as criterion variable, but no consistency in GPA predictability among the variables was obtained. For the 1968 group, however, the panel ratings in Bangkok proved the best predictor. Gue and Holdaway concluded that 'neither TOEFL nor the panels were outstandingly good as predictors', but their findings gave sufficient support for the continued use of objective tests of English proficiency as predictors of academic success and

for further experimentation in the use of interview panels.

2.5.3.9 In another study in the USA scores on TOEFL, the Graduate Record Examination (GRE) and GPA were obtained for a sample of 978 overseas students at 24 different graduate schools (Sharon, 1971). The mean TOEFL score was half a Standard Deviation above the mean of the reference group (TOEFL candidates 1964-69). The mean score on the verbal part (reading comprehension, vocabulary and verbal reasoning) of the GRE was more than one SD lower than the mean of the reference group (American students 1966-69) while the mean of the quantitative section (arithmetical reasoning, algebraic problems and interpretation of graphs, diagrams and description data) was a half SD above the mean of the American reference group. Correlations between GRE Verbal and GPA and between TOEFL and GPA were almost identical ranging from .21 to .41. When GRE was combined with TOEFL the correlations with GPA only increased to .23 to .42. A combination of TOEFL and GRE qualitative, however, gave correlations from .34 to .61 with GPA. Although these latter correlations were far from showing a very close relationship they showed an increase on the correlation with TOEFL only. Sharon concluded

(a) an overseas student may succeed in an academic programme - particularly at postgraduate level - with initial low verbal aptitude or ability in English and

(b) an English proficiency test may have greater predictive power when combined with another type of aptitude measure.

He also questioned the suitability of GPA as a criterion.

2.5.3.10 One of the features common to the data reported in many

of these studies is the high rate of success and the low rate of failure on the criterion of academic results as judged by either GPA in North America or by examination results in Britain. Although many students obtain very poor scores on language proficiency tests, they reach the criterion standard achievement after a time, often after having followed further English courses in Britain, Canada and the United States. This situation contributes further evidence to support Sharon's first conclusion above.

2.5.4 The major difficulty with criterion-related validity is the validity and suitability of the criterion. The reliability of teachers' ratings may be suspect unless the sample is small and the raters few in number and conversant with the rating procedure. Alternatively the sample may run to several hundred and involve a large number of raters. But the teacher rating method has been restricted almost entirely to the validation of proficiency tests in Britain, EPTB, ELBA and Chaplen (1970). Very few instances of using ratings are reported from North America. All the examples of use of teachers' ratings reported above have been for concurrent criterion-validation.

The use of other proficiency tests as criteria has been limited to concurrent validity studies. Since the validity of other tests is far from proven, such studies can do little more than indicate the extent to which the same results are yielded and the same general skills are being tested.

The English proficiency tests for which validity studies have been reported have been administered almost exclusively to students whose

native language is not English but who desire to go to Britain, Canada or the United States for further study through the medium of English. Consequently it has proved possible to adopt the criterion of academic achievement as exemplified in a final pass/fail result, or Grade Point Average, or similar achievement rating. This academic achievement criterion has been used in determining predictive validity. The pass/fail distinction is, however, very crude and the GPA is currently criticised as being an unsatisfactory method of assessing students' success, particularly at the graduate level.

2.5.5 The relationship betwen the criterion and the proficiency test is usually expressed as a product moment or point biserial correlation. The correlations reported are positive, frequently highly significant statistically, but generally in the order of .4 to .6. Higher correlations of around .8 have been reported mainly when the criterion variable has been another proficiency test administered concurrently or when it has been applied at a later date and when the sample to whom it is applied is following a course in the teaching of English as a Foreign Language. The correlations obtained nevertheless provide some evidence of the validity of the measure being investigated although such evidence cannot establish the validity of a measure beyond doubt.

Expectancy tables based on performance of subjects on both the criterion measure and the measure being validated appear to provide equally satisfactory sources of evidence of validity, particularly when the predictive power of a battery is being examined. This method of relating the measure to the criterion also provides more

information to the consumer of the test scores. The evidence reported above both in the form of correlations and expectancy tables from 1965 to 1975 seems to point to the suggestion that low proficiency in English, as measured by a proficiency test, may not be the barrier to successful study through the medium of English that it is generally thought to be.

2.6 Conclusions on the Validity of Proficiency Tests

A study of the validity of English language proficiency tests can only serve to underline the complexity of language proficiency testing and Carroll's view that validity is one of the 'persistent problems of foreign language testing' that will persist (Carroll, 1973 op cit).

On the basis of the discussion so far in this chapter we may conclude that

(a) very little is at present known about what constitutes language proficiency and consequently it is impossible to determine construct validity until more work has been done in this area;

(b) the distinction between construct and content validity in language testing is not always very marked, particularly for tests of general language proficiency;

(c) adequate sampling of one or more domains of discourse in the space of a very short time cannot be achieved except in cases of very restricted linguistic context;

(d) it is difficult to be certain that any criterion variable chosen is itself sufficiently valid.

2.6.1 It cannot be said with any certainty that any proficiency test is valid. Validation procedures such as those discussed above

yield evidence pointing to the validity of a language proficiency test as well as evidence suggesting lack of validity. When the different pieces of evidence are considered, it is possible to conclude that a particular test appears to possess some overall validity although it may be deficient in certain specific types of validity. This would seem to be the conclusion reached on the various English language proficiency tests cited above.

2.7 The Test Consumer's Viewpoint

This lack of certainty about the validity of different English language proficiency tests presents problems for the consumers of such tests, whether the candidate, the candidate's teacher, or an administrator concerned with the candidate's performance. The consumer is concerned with obtaining the answers to two major questions, which constitute the layman's desire for evidence of validity (see para 2.2.2 above). These questions are:

(a) Is the test a reasonable test of English proficiency? and

(b) To what extent does the test suit my/our purpose?

The first question relates to internal validity and the second to criterion-related validity. But obtaining answers is made more difficult because of two factors.

2.7.1 The first, discussed in the preceding paragraphs, is the complexity of establishing absolute validity and the apparent impossibility of doing so in the light of current linguistic and psychometric theory. One solution to the problem would be to abandon the search for absolute validity and rely on intuitive judgements of construct, content and face validity. However, past

reliance on intuition and the paucity of documented validity studies have led consumers to doubt the validity of many existing language tests. Another solution would be to intensify the search for validity through further experimentation and follow up studies, thus building up a body of information on the performance of tests leading in the first instance to more satisfactory interpretive information and secondly to possible modifications to the test design.

2.7.2 The second factor is the increase in the numbers of candidates taking English language proficiency tests. 109,366 foreign students are reported to have taken TOEFL from February 1964 to June 1969 (TOEFL, 1970) and 473,944 foreign students took TOEFL during a shorter period from September 1978 to August 1980 (TOEFL, 1981). The total number of overseas students in Britain increased by more than 35% in 3 years from 69,283 in 1970/71 to 95,209 in 1973/74 (British Council, 1975) and by a further 25% to 119,559 in 1978/79 (British Council, 1980), although not all these students were tested in English. The number of candidates for the Cambridge Proficiency Examinations has been increasing rapidly (UCLES, 1973 and 1980) and the number of candidates tested overseas by the British Council increased dramatically from 1971 to 1973 (Hindmarsh, 1977: 24) and has increased steadily since. In addition to the obvious problem of following up large numbers of candidates, certain characteristics of the candidate population may change gradually over a period of years. The majority of candidates for English language proficiency tests are preparing to attend various types of academic and practical courses in an English speaking country. With the rapid development of secondary

and tertiary education in most countries outside Europe and North America, with recent economic changes brought about by the rise in oil prices, and with recent increases in tuition fees in Britain and Australia, taking courses in an English speaking country is now open to candidates who are more likely to be nearer the average in terms of intellectual ability and academic achievement than their more 'elite' predecessors. The purposes for which they need English may vary greatly. English is no longer required for academic purposes only but increasingly for occupational purposes, which frequently means less need for reading and writing proficiency but a greater dependence on oral/aural proficiency, perhaps within a fairly limited range of contexts.

2.7.3 The test consumer is thus left with a limited number of English proficiency tests about whose validity and suitability he cannot be at all sure. The individual consumer, eg the candidate or the teacher, is not in a position to carry out any form of research whereas the institutional consumer may well be dealing with large numbers of students which cannot be handled. A near impasse has been reached where the consumer is unable to carry out criterion validity studies for the above reasons and where the test constructor is equally powerless to go beyond establishing some degree of internal validity because of lack of time or lack of access to students who have taken the test. This unsatisfactory situation suggests that both test constructors and consumers should come together and examine the question of English proficiency test validity afresh.

2.8 An Approach to Test Validation

Validation would appear to be a constant process in the life of any proficiency test, to be carried out in two stages - at the construction and the performance stage. The two characteristics being looked for would be

(a) consistency of construction

(b) acceptability of performance.

The underlying rationale of this approach is that a test should be evalauated according to its own internal criteria of construct and purpose and not according to some externally imposed criteria.

2.8.1 Thus, the first step in this approach would be to establish the theory or construct of language proficiency on which the author had based his test and check that the test had been constructed in accordance with those principles. This would entail assessing the appropriateness of the item types as well as of the language sampling and taking account of constraints imposed, such as time available for the test, sound reproduction equipment available, frequency of administration and whether marking facilities were satisfactory. The theory of the construct would not be evaluated at this stage but would be accepted as being 'reasonable' in the light of trends in applied linguistic research. Dealing with construct and content validity in this way means that in effect a proficiency test is being treated as an ordinary achievement test based on a given syllabus or set of materials, but with the 'syllabus' determined by the author(s) of the test. This relates to the reasoning of Cartier and Petersen (1975, op cit) whereby they advocated that Defense Language Institute proficiency tests should be based on the content of the Institute courses.

2.8.2 Once the construction of the test was accepted as being consistent with the principles on which it was based, the next step would be to assess its performance in the light of its stated purpose.

A proficiency test may aim to reveal whether candidates have attained certain levels of proficiency which have been determined with or without reference to a learning group. It may aim to indicate whether a candidate has reached a level which would enable him or her to carry out certain tasks in the target language. The performance of the test would then be assessed by comparing the performance of candidates on the test with their performance according to another appropriate criterion and determining the acceptability of the results obtained.

This is likely to be achieved in two stages. The first would involve monitoring the test's performance while it was being piloted and pre-tested, and the second would entail assessing the test's progressive performance over a period of time. Initial assessment may well have to be limited to concurrent validity studies and serve to establish interpretive criteria. Assessment of progressive performance using other criteria may serve to confirm or modify the interpretive criteria initially established.

2.8.2.1 Reliability is a characteristic which should underlie the performance of a test. It would be important to establish maximum reliability at the construction stage and to check the test's continuing reliability. The degree of reliability would be determined by the construct and the purpose of the test. Language

proficiency tests should ideally contain tests of production, but the language produced, whether in written or spoken form, cannot be anticipated with precision. What is said or written on a particular occasion in a certain context may well be unique, and may well represent something that is very personal or special to the author. In such cases commonly adopted norms of reliability might have to be changed.

2.8.3 An attempt to establish the basic validity of a language proficiency test should therefore take account of:

(a) the consistency of the construction with the theory of language proficiency on which it is based, regardless of whether the theory is entirely acceptable to current linguistic thinking, and

(b) the relationship between performance on the test and performance on one or more appropriate criterion measures by the initial group and by subsequent groups.

2.8.3.1 The major emphasis should thus be on criterion-related validation procedures, and particularly on those relating to predictive validity. An English proficiency test should yield information on the levels of English proficiency attained by candidates and also on the adequacy, or inadequacy, of those levels in respect of the use to which the candidates expect to put this English (see para 2.8.1 above). The criterion measures should yield information on the levels of English proficiency attained by the candidates once they are engaged in their anticipated activities and on the adequacy, or inadequacy, of their English proficiency

for those activities. Since there will normally be a certain lapse of time between the administration of both sets of proficiency measures, and because of other intervening variables, it would not be appropriate to expect candidates to attain the same levels of proficiency (either on the same scale or on an absolute scale) in the criterion measures. However, the criterion measures should indicate the extent to which the levels now attained are adequate or not. In other words the performance on criterion measures should confirm the general expectations or predictions made in respect of each candidate on the original measures. While the expected performance may be measured initially on a delicate scale, it will be sufficient to establish in a more general way whether the expected level of adequacy has been attained or not.

2.8.3.2 The major practical implication of this approach is that validation would become an on-going feature of proficiency testing. With this approach the performance by the initial group tested would be important, but not as 'all important' as at present. Other groups could periodically be subjected to the same or refined criterion measures to check on the validity of the test. This could yield modified and more meaningful interpretive information about scores obtained which would be a major advantage for test consumers. The theoretical implication of this approach would be that feedback from such studies could lead to modification of the construct of the test and so to further experimentation aimed at improving the construct validity. A further general advantage would be that the test would be periodically reassessed in the light of new linguistic insights and new, or changing, needs of the consumers.

2.8.4 In order to test the validity of this approach, a number of proficiency tests of differing types would have to be investigated, and follow up studies carried out. The simultaneous investigation of a number of proficiency tests would present formidable logistic difficulties, but the studies could be spread over a longer period by taking one or two tests at a time. In this way a validation procedure could be tried out and modified in the light of any unanticipated flaws or omissions before being applied to the next set of tests.

2.8.5 In view of the need for further evidence of criterion-related validity for English language proficiency tests, it was decided that the approach to test validation outlined in the preceding paragraphs should be investigated. While trying out the approach it would also be possible to gather new information on the validity of existing tests which would be of benefit to the consumers of the tests. It was thought that by the end of the study a framework for future validation, with special emphasis on establishing predictive validity, could emerge which would prove logistically feasible and which constructors and consumers of future English proficiency tests could utilise.

2.8.6 Two measures were chosen for investigation - the English Proficiency Test Battery, Short Version, (Davies, 1964), hereafter referred to as EPTB, and the subjective assessment procedures used by the British Council in many of their offices overseas, hereafter referred to as BCSA. The details of BCSA have not been published, but are explained in a Standing Instruction issued internally by

the British Council (1972).

The measures were chosen for the following reasons. Firstly, they represent two very different approaches to proficiency testing. BCSA is a traditional and completely subjective form of assessment relying heavily on the ability of each examiner to choose appropriate tasks for the candidates and to assess their performance satisfactorily. EPTB, on the other hand, is entirely objective and a product of the psychometric-structuralist trend in language testing. Secondly, both tests were widely used by the British Council for a number of years to assess the adequacy of the proficiency in English of students wishing to enrol in courses of study in Britain or other English speaking countries. The results of the assessments were extremely important both for the candidates and for the institutions to which they applied, since they affected the lives and careers of the former and the enrolment policy of the latter. The context, therefore, was not only pertinent but part of everyday life and in no way hypothetical. Thirdly, the data necessary for such a study was more accessible than data for other tests.

2.9 Formulation of Hypotheses

In order to carry out the study it was necessary to establish a set of hypotheses based on the approach to test validation as summarised in paragraph 2.8.3 above and relating specifically to the two proficiency test measures chosen for investigation.

2.9.1 First Hypothesis: The construction of the English Proficiency

Test Battery, Short Version (EPTB) is consistent with the theory on which it is based, and is compatible with current thinking in linguistics and language testing.

2.9.2 Second Hypothesis: The construction and administration of British Council Subjective Assessment procedures (BCSA) are consistent with the theory on which they are based, and are compatible with current thinking in linguistics and language testing.

2.9.3 Third Hypothesis: There is a significant positive relationship between performance on EPTB by an appropriate sample of students who are non-native speakers of English and performance on appropriate criterion measures.

2.9.4 Fourth Hypothesis: There is a significant positive relationship between performance on BCSA by an appropriate sample of students who are non-native speakers of English and performance on appropriate criterion measures.

2.9.5 Fifth Hypothesis: There is a significant positive relationship between performance on EPTB or BCSA and performance on appropriate criterion measures for experimental groups affected by such intervening variables as attendance at remedial English courses, the type of course being followed, the general subject area of study and country of origin.

2.10 Stages of the Investigation

It was decided that investigation of the hypotheses should be carried

out in the following stages:

(a) assessment of the extent to which the construction of each set of tests may be said to be consistent with their underlying theory, and with the current theory of language proficiency testing,

(b) selection of an appropriate sample of subjects who had recently taken the tests, and procurement of their scores or grades,

(c) evolving new criterion measures to administer to the sample once they had taken up their studies in Britain, and

(d) analysis of the data obtained and the drawing of conclusions on the relationships between the sample's performance on the initial measures and on the criterion measures.

CHAPTER 3 ENGLISH PROFICIENCY MEASURES INVESTIGATED

3.0 Introduction

In this chapter the measures to be investigated, the English Proficiency Test Battery (EPTB) and the British Council Subjective Assessment (BCSA), are described and then assessed in the light of the theory on which they were based, and in the light of current thinking - see para 2.10 above.

3.1 Design of EPTB

The design of the original English Proficiency Test Battery was drawn up on the basis of certain linguistic categories to be tested and on the psychometric principle of work sampling. The desirability of testing oral and written expression was recognised at the outset. But because of the practical difficulties of administering tests of language production and the theoretical problem of the reliability of such tests, the decision was taken to concentrate on testing through listening and reading (Davies, 1965: 59-62).

3.1.1 Some of the tests were based on the linguistic levels of phonology and grammar, since these were deemed to be of importance for non-native speakers of English who intended to embark on academic studies through the medium of English. They were

Test 1	Listening	Phonemic discrimination - words in isolation
Test 2	Listening	Phonemic discrimination - words in sentences
Test 3	Listening	Intonation and stress discrimination
Test 7	Reading	Grammar - recognition of appropriate features.

All these tests contained discrete point items which could be scored objectively. Because of the problem of adequate sampling there were no tests of vocabulary.

3.1.2 The remaining tests in the original battery were designed to be representative work samples and included

Test 4	Listening	Comprehension
Test 5	Reading	Speed reading
Test 6	Reading	Comprehension

Although each item in these tests is separate and objectively scored, the correct answers or choices are in most cases dependent in part on satisfactory comprehension of the meaning of the texts as a whole and in part on satisfactory completion of adjacent items. Few items are totally independent. The highest possible reliability was aimed at for all the subtests.

3.1.3 The battery was tried out on three different groups:

- overseas students studying in Britain, in their first year where possible (N = 496)
- overseas students studying in their home countries (N = 238)
- native English-speaking students (N = 267)

On completion of the try-out item analysis, computation of reliability, factor analysis, and concurrent and predictive validity studies were carried out.

3.1.4 Following the try-out, and in response to operational needs of the British Council and the NFER, a Short Version was produced. Of the 7 subtests listed above, two were excluded from the Short Version and one was made optional. The listening comprehension test (Test 4) was excluded. It consisted of comprehension questions based on lecturettes but yielded very few items. Reliability was consequently very low at .51. Moreover, factor analysis suggested

that in addition to being associated with a textual listening factor, it was also associated with a general reading comprehension factor. Test 2, phonemic discrimination of words in sentences, was also excluded since it was represented factorially by Test 1, which had slightly better reliability (Davies 1965: 164). Test 5, the test of reading speed, was retained as an optional Part II of the Short Version, while the remaining 4 tests were retained as Part I of the Short Version. The tests were renumbered and the Short Version was made up as follows:

Part I	Test 1	Phonemic discrimination
	Test 2	Intonation
	Test 3	Reading comprehension
	Test 4	Grammar
Part II	Test 5	Reading speed

It is consequently this Short Version which is investigated in this study. Short Version, Form A, was used first by the British Council to assess the English proficiency of overseas students applying for courses of study or training in Britain and other English-speaking countries in 1965. As numbers of candidates, and consequently security risks, increased, parallel forms of the Short Version were required. Three parallel forms were subsequently produced - Form B (Davies, 1965a), Form C (Davies and Moller, 1973) and Form D (Davies and Alderson, 1977). Form D incorporated changes in the listening sections. Listening tasks 1 and 2 were replaced by a listening comprehension test in three parts:

- identification of sentence stress	25 items
- comprehension of discourse	25 items
- comprehension of authentic interview	17 items

The various forms of EPTB were in use by the British Council until 1980.

3.1.5 Standardised scores are reported for each of Tests 1 to 4. Each test has a mean of 10 and SD of 2. A total score - the sum of the four subtest scores - is reported for Part 1, the mean being 40 and SD 6. Part II scores are not standardised, but the mean for Form A is 70 and SD 34.

Expectancy tables for Form A showed an optimum cut off against the predictive criterion of 34 and an optimum concurrent cut off at 36 (Davies, 1965: 216-222). Over 80% of the subjects gaining these scores or above were rated as successful either in final examinations or by their tutors. However the notion that students gaining a certain score and above would be successful and that those gaining one mark or a few tenths of a mark below that score would not, is manifestly wrong, though convenient administratively. In order to allow for the fact that scores are simply points on a continuum, and to allow for standard error of measurement, the banding system was introduced for Forms C and D. The interpretation of the bands of scores is given in Table 3.1. Clearly this approach eliminates the type of interpretation that assumes that candidates with, for example, 39.8 have insufficient English and that those with 40.2 have sufficient English. But there is still the problem of how to interpret scores that are within about two standard points (or one third of a standard deviation) above or below the cut offs at 34.0 and 40.0. It is at this point that extra information, such as performance on Part II or on any supplementary essay or oral tests, contributes to the interpretation.

Table 3.1 Interpretation of scores: EPTB Short Version, Forms C and D, Part I

Below 34.0	- insufficient English to follow a course. A minimum of 6 months full-time English tuition will be needed.
34.0-39.9	- candidate will probably need some preliminary intensive tuition to improve his abilities in English. The period of tuition may vary from 4 to 24 weeks.
40.0 and over	- should have sufficient English to follow a course in his subject in Britain.

3.1.6 A more detailed examination of the subtests which comprise the Short Version will show the extent to which the test is constructed in accordance with its own principles and with those of its time.

3.1.6.1 The first two subtests are tests of listening with a linguistic basis. Test 1 is a test of phonemic discrimination which samples vowel and consonant contrasts in initial, final and medial positions. A sample item is[1]

subject hears — tʌn tʌn tɜn

subject reads codes — AB BC ABC AC 0

subject selects appropriate code and circles it[2].

- correct response: AB ie first two words sound the same. There are 58 items.

3.1.6.2 Test 2, a test of intonation and stress discrimination, consists of short dialogues contrasting intonation patterns. The

---

1. All sample items are taken from EPTB Short Version, Form A.

2. The code has been changed to 12 23 123 13 0 in Forms C and D.

example below illustrates the contrast between the rise-fall and the fall on hungry and the contrast between the fall-rise and the the fall on eat and anything. In discourse in which factual or neutral information was being stated, all these words would convey the tonic and subsequent falling tone. This is the most frequent intonation pattern and the one that the listener would expect unless the context indicated otherwise. However, in this utterance the intonational patterns on these final words vary to convey special meaning and attitudes. Sample item:

Subject hears twice: (John) Let's have a meal. Not that you are hungry.

(Mary) I don't want to eat? What an idea. Of course I can't eat anything.

Subject reads
(a) Mary wants to eat
(b) Mary can eat nothing

The subject is required to put a tick beside each statement that he thinks is correct. In this example he should tick statement (a). As Davies points out (1965: 92-3)

> 'Here the important contrast is shown by the falling-rising nucleus on anything, implying that Mary likes some but not all things.'

Similarly the rise-fall on hungry is intended to convey some irony, and the fall-rise on eat indicates that Mary means that she does want to eat. There are 22 short dialogues, each followed by two statements which must be judged true or false by the candidate. For technical reasons six responses are not considered, and so the test yields 38 items.

3.1.6.3 The third linguistic test, Test 4, is a test of grammar. Subjects have 47 unrelated sentences to read. A part of each

sentence contains three alternative words, or phrases from which the subjects have to select the one that they think a native speaker of English would use, as in the following example.

1. Do you like
2. Would you like to come to tea tomorrow?
3. Could you like

Correct response is 2.

The point of grammar being tested here is the appropriate modal to be used in the context of a polite question (really an invitation) using the verb <u>like</u>. The other items test different grammatical features, and the subtest is to be completed within 15 minutes, thus allowing just under 20 seconds per item.

3.1.6.4 The two remaining subtests are of the work sample type. Test 3 is called a test of reading comprehension and consists of two short passages with certain words deleted, except for their initial letters, as in the example below. Subjects have to complete the deleted words by writing them in the spaces on the test paper. 49 words have to be completed within 5 minutes - a rate of one word every 6 seconds. The rate of deletion is about one word in every three. Example:

> B..... changes i..... t..... home are less revolutionary, a..... easier t..... assimilate, t..... changes i..... industry.
>
> The words to be completed are But, in, the, and, to, than, in.

As a work sample this kind of task clearly corresponds to reading under pressure and can only be completed satisfactorily if first the general context of the passage is grasped, and secondly, if is realised that only function words have been deleted - a fact

that the candidate has to work out. While this subtest could not be termed as a purely 'linguistic' test, its linguistic content is very important. The principal linguistic level being tested is that of grammar, and the concept underlying this test is redundancy. A native speaker of English should be able to supply the missing items without difficulty, whereas a non-native speaker may need more clues than the mutilated text provides.

3.1.6.5 The fifth (optional) test is a test of reading speed. Subjects are required to read as much of a text of one thousand words as they can in 10 minutes. But about twenty percent of the words have been inserted into the original text. These words (known technically as intrusive words) are inappropriate and impede comprehension. The subjects must underline those words that they consider do not belong to the test. Example

> 'Our British policy for speak higher education is tenable girl on certain assumptions. The first did assumption is....

The first two intrusive words are underlined as examples, but subjects have then to identify did as not belonging and underline it. This test is not dependent on any one linguistic element and is the most integrative of the battery.

3.1.7 Reported reliabilities are given in Table 3.2 (British Council, unpublished, 1973). It is to be noted that although reliabilities for Form C, except for Test 2, are lower than those originally reported for Form A, reliability based on the 1973 sample's performance on both forms is higher for C than for A. The discrepancy between the two sets of figures for Form A is

probably due to the difference in the size of the two samples and the probable greater homogeneity of the 1973 sample.

Table 3.2 Reliability of EPTB, Forms A and C

Test	No of items	1964 Form A	1964 Form C	1973 Form A
1	58	.91	.85	.82
2	38	.75	.75	.62
3	49	.94	.90	.93
4	47	.89	.85	.81
5	196(A) 163(C)	.97	.96	.96

1964 sample, N = 496 overseas students in Britain - split half corrected
1973 sample, N = 167 overseas students in Britain - KR 21 formula

However, the high reliability of subtests 3 and 5 is established across Forms and samples, and all sets of figures place the reliabilities of the subtests in the same descending rank order, viz: Test 5, Test 3, Test 1, Test 4 and Test 2. The low reliability of Test 2 is outstanding in all three sets of figures and must cast serious doubts over the validity of the intonation test.

3.1.8 Davies (1965: 190-1) has claimed validity for the battery, considering the reliability adequate and the overall concurrent and predictive validity coefficients of over .45 to be satisfactory. Tutors' rankings on a 5 point scale were used as the concurrent criterion and grades or examination results at the end of the academic year as the predictive criterion.

3.2 Consistency of EPTB with the Theory

It is now possible to consider the first hypothesis stated at para 2.9.1 above, namely the extent to which EPTB may be said to be consistent with

the theory underlying the test.

3.2.1 Davies (1965: 52) specified three essential needs of a language test - 'linguistic content; language control; desired validity' - and these formed the basic concepts of the theory underlying the test battery.

The 'linguistic tests' were designed to satisfy the need for linguistic content, and the levels of language tested were the phonological and syntactic, as exemplified in Tests 1, 2 and 4. Language control was tested in the work sample tests through general reading comprehension (Test 3) and reading speed (Test 5). Textual listening comprehension was included in the original battery but was dropped from the Short Version - see para 3.1.4 above. The need for validity was met by establishing two sets of criteria - one concurrent and one predictive.

3.2.2 The linguistic level of phonology was divided into segmental and supra-segmental features, as examplified in Tests 1 and 2 respectively. Test 1 would appear to be entirely consistent with the theory. It sampled the vowel and consonant phonemes widely with emphasis on problem phonemes. There is, however, evidence that this test is relatively easy (see the means recorded for performance in Test 1 in Table 3.3 below). It is also debatable whether success in Test 1 really denotes ability in English. It would appear to denote some language aptitude, since the learner has to recognise and differentiate between sounds without reference to context or meaning. The discarded phonemic discrimination tests with the words

in sentences, though more difficult to construct, would have had greater face validity and had the advantage of greater predictive validity (Davies 1965: 169). However, it is doubtful whether it would have increased the validity of the battery.

Testing intonation is a challenge to the test constructor. In order to effect changes of meaning, emphasis, attitude and mood, sometimes only the slightest changes in intonation are necessary. The dividing line between one meaning or attitude and another is therefore often very difficult for the native speaker to identify. Because of these fine distinctions constructing items and reading them for non-native speakers of English is a most exacting, if not impossible, task. It is noteworthy that of the seven major proficiency tests listed in Table 1.1, the testing of intonation is attempted in only two, EPTB and ELBA. ELBA contains only ten items in simple sentences, whereas EPTB has 38 items in short dialogue.

However, the EPTB Test 2 is not solely a test of intonation but a test of conversation comprehension as well. Davies was aware of the weakness of the test (1965: 90-91). He comments firstly that 'several of the items were known to contain non-intonational clues (eg lexical ones) which contaminated the signal it was desired to isolate', and secondly he admits it is a very 'ad hoc way of going about testing conversation from a linguistic point of view in the absence of an acceptable and accepted inventory of the intonational patterns'. Chaplen (1970: Appendix 2, xvi) comments on the uncommonly heavy communication load which many of the intonation patterns have to carry and the consequent unnaturalness of the language in many of the items. He also notes that since

the subject has only to make a yes/no choice for each item, 'chance success due to guessing must give rise to a considerable error factor'. In addition, the reliability of this test is the lowest of all the subtests and cannot be considered satisfactory. It is clear that a serious attempt was made by Davies to construct a test of intonation consistent with the theory behind the battery but his choice of test format did not make for high reliability and has led to his producing samples of English of some ambiguity and of suspect authenticity. This is the least satisfactory and least valid subtest of the Short Version in spite of previous claims made for its validity. It is to be noted that this test has been removed from Form D and a test of sentence stress substituted.

The third linguistic level test is a test of grammar (Test 4). It is a discrete item test in multiple choice format which has become a basic subtest type in language proficiency tests. In Table 1.1 only the French CGM 62 test fails to include such a subtest. This type of test has become a modern classic and generally shows high reliability. The most important feature of this kind of test is the sampling of the grammar tested. In the case of EPTB half of the items are devoted to verb forms, verb tenses and phrasal verbs (including modals). It is deficient in items on article usage and prepositions, although these do appear in the 'common errors' category. Increasing the total number of items to 60 would improve sampling, increase reliability and only add an extra five minutes for administration.

3.2.3 Language control was the basis of the two remaining tests of the Short Version, both of the work sample type. Whereas items

in the three linguistic tests discussed above were constructed around a corpus of discrete linguistic features, the two work sample reading tests were based on running texts chosen for their appropriateness to intending academic students. Both tests are timed tests and candidates are obliged to work under greater pressure than in the linguistic tests. Furthermore, administration of them is very rapid, and they represent a very efficient use of the time available. Test 3 lasts only 5 minutes - just 10% of the time required for Part I. It yields 25% of the items and total score for Part I. Test 5, while optional, lasts only 10 minutes, or one sixth of the time required to administer the whole Battery. The A and B forms of the Battery contain as many items in Test 5 as the whole of Part I. (Forms C and D contain about 17% fewer items.) As shown in Table 3.2 above, these subtests have been shown to be consistently the most reliable of the subtests.

Test 3 is ostensibly a test of reading comprehension in modified cloze format. The skill of reading is the dominant language skill the candidate is required to exercise, and the writing required for completion of the items is kept to a minimum. But the items themselves are all function words, and the test could be considered another test of grammar with items set in context. While the general task can be considered to be consistent with a work sample, this subtest must still be considered in part a test of the linguistic level of grammar. This being so, Test 3 cannot be said to be entirely consistent with the stated basis of the battery. Nevertheless it is reasonable to include a subtest which is neither wholly linguistic nor wholly work sample.

The reading speed test shares one main feature with the comprehension test - the text has been modified or mutilated. Whereas the text of test 3 had had part of words deleted, the text in Test 5 had had words added to it. But these added words do not represent any one linguistic level, since they have been taken from another running text. Thus the candidate has to draw upon both his knowledge of the semantics of English as well as of the grammar. This subtest cannot therefore be said to be a test of the linguistic levels of English, but is certainly a difficult test of language control. There is, however, one aspect of this test which is not entirely satisfactory. Two different types of word are added - first Welsh words (presumed to be unfamiliar to all candidates) and then English words. Although the difference in the words is usually indicated in the instructions, 'some foreign and some irrelevant English words', the test can be considered two different tests since there are two distinct, though similar, tasks. In Form A 44 (22%) of the items are Welsh words, and this part of the test has proved easier. (No results have been published to substantiate this, but it is the writer's experience of marking this subtest that most candidates have identified the Welsh words with at least 80% success and that the speeding and the English words provide the difficulty and the discrimination.) In Forms C and D the number of Welsh words has been reduced proportionately to 36.

It has already been noted above that the work sample tests have consistently proved the most reliable of the Short Version. They have also been the two most difficult tests, as the facility values

Table 3.3 Comparative facility values for EPTB: Form A (1964), Form B (1965) and Form C (1973)

Forms	A %	B %	C %
Test 1	72	70	75
Test 2	56	59	60
Test 3	56	54	57
Test 4	71	69	76
Test 5	35	N/A	38

(Values are raw score means expressed as %)

obtained by the try-out groups for Forms A, B and C, set out in Table 3.3, show. In addition Test 3 has proved the most discriminating of the subtests, as shown by the raw score Standard Deviations expressed as a percentage of the total number of items in each subject in Table 3.4.

Table 3.4 Comparative SD for EPTB Form A, Form B and Form C

Forms	A %	B %	C %
Test 1	17	19	14
Test 2	14	16	15
Test 3	27	21	21
Test 4	18	19	15
Test 5	17	N/A	17

(Values are expressed as % total items in each subtest)
Source: EPTB Form C Scoring Instructions, 'Comparative Statistics' (unpublished)

3.2.4 The third element in Davies' basic theory for the EPTB was validity. By this he meant criterion validity. The elements of linguistic level and language control could be said to account for content and, more particularly, construct validity, and these elements have been discussed in the present chapter. The adequacy of

the criterion validity of the battery is the subject of the present investigation, reported in the following chapters, and is basic to the examination of the third hypothesis (see paragraph 2.9.3).

3.2.5 The detailed examination of the subtests in the preceding paragraphs leads to the conclusion that for the most part the subtests of EPTB, as exemplified in Form A, are consistent with the theory underlying the test. The basic design is consistent. The three linguistic tests systematically test knowledge of the levels of phonology - segmental and suprasegmental - and grammar. It has been noted, however, that it is not necessarily clear whether success in Test 1, the test of phonemic discrimination, denotes real ability in English, and so there is doubt about its appropriateness. It has been further noted that the test of intonation is very difficult to construct and that this subtest has certain limitations - the authenticity of some of the utterances is questionable, the tasks in the test demand some evidence of language control as well, and the reliability is low. Further sampling of structural items is desirable in the grammar subtest. This weakness is partly offset by the fact that subtest 3 cannot be considered solely a test of reading comprehension. It is certainly in part a further test of grammar, and so not entirely consistent with the notion of a test of language control. The test of reading speed can be considered consistent.

3.2.5.1 But the fact that not all the tests can be categorised as wholly 'linguistic' or wholly 'language control' does not militate against the consistency of the battery as a whole. It

can be concluded that the battery consists of

- two purely 'linguistic level' subtests (phonemic discrimination and grammar)

- two subtests which are partly 'linguistic' and partly 'language control' (intonation and reading comprehension), and

- one 'language control' subtest.

One further point should be noted. A test constructor cannot be responsible for the way in which his test is finally used. It was originally intended that all the subtests of the Short Version should be administered. But it has become common practice to exclude the reading speed test from most administrations. Consequently the test users have unwittingly compromised the basic design of the battery and upset the already uneasy balance between linguistic level and language control, so that the battery being administered is predominantly one of linguistic level tests.

3.2.6 The greater emphasis on linguistic levels was in keeping with testing theory and practice during the psychometric-structuralist trend, which was particularly strong in the early 1960s. It is not in keeping, however, with current theory of language testing. It has been noted above (para 1.3.1) that one of the major characteristics of the latest trend in language testing has been the concern with testing communication. In her preface to the second edition of her book Valette (1977: v) notes that the 'growing interest in language as a means of interpersonal communication has led to the development of a variety of tests of communicative competence'. However, it has been pointed out that Valette has not given examples of published standardised tests of

communicative competence (Davies, 1978: 149), since none exist. Attempts are currently being made to produce such tests. One such attempt is being undertaken by the British Council. The aim is to create a test system 'that will give an answer to the question as to whether a candidate is likely to be able to meet the communication needs of a given course of study' (B J Carroll, 1978: 4).

EPTB cannot be said to be able to provide an answer to that question in respect of individual candidates and is not in harmony with the current search for valid tests of communicative competence. However, Davies, in his discussion of testing the communication skills (Davies, 1978: 221-6), points to the dilemma that is bound to face the constructors of a standardised test of communication. On the one hand specifications of the types of communication will be required, but on the other hand, as Morrow (1977: 23) has pointed out, one of the essential characteristics of the use of language in communication is unpredictability! And that this cannot be specified, except in the most general terms.

3.2.7 The foregoing discussion leads to the conclusion that the construction of EPTB, Short Version, is consistent with the theory on which it is based, with some shortcomings arising from low reliability and some imbalance between linguistic and language control subtests. It is not, however, compatible with current thinking on linguistics and language testing, as outlined in Chapter 1 and in the paragraphs above. There is, however, no evidence to date that more recent tests with their emphasis on functional proficiency and communication have given significantly more accurate

information on the English of non-native speakers of English.

3.3 Design of BCSA

It is now appropriate to consider the hypothesis at paragraph 2.9.2, namely that British Council Subjective Assessments (BCSA) are consistent with the theory on which they are based.

3.3.1 Historically the BCSA developed because of the need to provide an assurance to British colleges and other institutions that certain students being sent from overseas countries had adequate English for the course to be followed in Britain. The only British examination or test in English as a Foreign Language to be administered outside Britain before EPTB was the Cambridge Proficiency in English. But that examination was only administered twice a year and the time between sending in an application to sit the exam and obtaining the result overseas could amount to as much as six months. In addition to the early application required, at least one paper involved students following a special course of study (Hindmarsh, 1977: 22). A set of procedures was thus required that could be administered rapidly and which could give an on the spot indication of a student's adequacy, or lack of it, in English.

The need for these assessments arose in the 1960s, a period when the traditional or pre-scientific trend in language testing was current, and before the major psychometric-structuralist tests of English had been produced. Informal tests of English were already being conducted on a small scale in British Council offices overseas, but these became formalised and more structured. In 1960

the British Council issued Overseas Standing Instructions on the assesment of proficiency in English indicating in general terms how subjective assessments were to be carried out overseas. When EPTB was finally ready for use in 1964/65 it was still not possible to use the battery in all countries and so the subjective assessments continued to be given, and in increasing numbers. In 1972 a new Overseas Standing Instruction (OSI 2/72) was issued by the British Council incorporating revised general instructions for the administration of both the EPTB and the subjective assessments. No major revisions affecting the substance of the assessments have been made in these instructions between 1972 and the phasing out of ECSA in 1980.

3.3.2 It is hardly appropriate to talk of the design of the ECSA since the procedures evolved on an ad hoc basis over the years. However, the document OSI/72 does contain sufficient instructions, suggestions and examples for the framework to be set down and the procedures described.

'The applicant's likely performance in Britain is what is being assessed. This is not necessarily the same as his performance in English at home.' The assessor must not just evaluate the performance of the candidates on the date of the test but must estimate what their performance is likely to be at a later date in the native-speaking environment. In addition, the assessor may recommend a period of tuition in English on candidates' arrival in Britain, although he is not bound to do so. The aim of the assessment is to determine

'Whether the candidate will be able to participate profitably in a British university course or practical training attachment.

Will he be able to understand what he hears (not always the same thing as what is said to him directly), and will English people be able to understand what he says?

Can he read and write English without undue slowness or difficulty?

The above are the essential questions which lie behind the very careful wording of the various categories in the grading accompanying the subjective test form.'

(Quoted from Appendix C of the British Council Overseas Standing Instruction 'Assessments of Proficiency in English, London 1972.)

In addition to the above instructions and suggestions the assessor is asked to give one grade on the scale A to E, and in accord with the definitions laid down (see Table 3.5) for each of the four abilities listed, without using plus or minus signs. No overall grade is to be given. The abilities are

1. understanding of spoken English
2. ability to speak English
3. understanding of written English
4. ability to write English

3.3.3 No materials are given. Suggestions are made, and in some cases sample texts are given, or sample responses with comment on their evaluation are given. It is suggested that candidates should be asked to write first. This should take 15 to 20 minutes and could be a narrative based on a picture sequence. A sample picture sequence is given along with twelve sample pieces which were actually written on the picture sequence under test conditions.

Table 3.5 BCSA grades of proficiency in English

1. UNDERSTANDING OF SPOKEN ENGLISH

A When addressed in normal to fast English with no concessions made to the fact that he is a non-native listener, and with only very occasional rephrasing or repetition, he understands everything.

B When addressed at normal or slightly less than normal speed, with a few simplifications of expression, and occasional rephrasing and repetition, he understands almost everything.

C When addressed at less than normal speed, in mainly simple sentences, with some rephrasing and repetition, understands almost everything.

D When addressed rather slowly in only simple sentences with frequent repetition and rephrasing, understands almost everything.

E Even when addressed slowly, and with frequent repetition and rephrasing, understands little or nothing.

2. ABILITY TO SPEAK ENGLISH

A Approaches native-speaker competence in accuracy, clarity and range of expression. With little or no effort on the part of the listener, he is fully intelligible.

B Speaks fairly fluently, with some inaccuracy in pronunciation and some restriction in ability to handle complex structures and vocabulary. With some effort on the part of the listener he is almost fully intelligible, but his range is limited.

C Speaks haltingly in mainly simple sentences with a considerable amount of inaccuracy in pronunciation and/or grammar. With a good deal of effort on the part of the listener and some requests for repetition he is largely intelligible, but his range is very limited.

D Speaks haltingly in simple sentences only, with considerable inaccuracy in pronunciation and grammar but with evidence of ability to use basic structures. With a good deal of effort on the part of the listener and some requests for repetition most of what he says is intelligible, but his range is extremely narrow.

E Even with a good deal of effort on the part of the listener, and frequent requests for repetition, he shows little or no ability to express himself.

3. UNDERSTANDING OF WRITTEN ENGLISH

A Reads fairly difficult English on general subjects at normal speed with near-native comprehension. He may have to use the dictionary occasionally.

B Reads fairly difficult English rather slowly, but with general comprehension. Reads straightforward narrative or exposition at reasonable speed with almost total comprehension.

C Reads fairly difficult English rather slowly with extremely limited comprehension. Can follow the outline of a straightforward narrative or exposition, though misses much of the detail.

D Reads straightforward narrative or exposition very slowly and manages to follow only the bare outline of the piece.

E Reads straightforward narrative or exposition with little or no understanding of its content.

4. ABILITY TO WRITE ENGLISH

A Approaches writing competence of a fully literate native speaker in accuracy, range and clarity of expression but may produce very occasional errors. Without any effort on the part of the reader, he is readily understood.

B Is generally able to use complex structures but is more stilted and restricted than the native speaker, and makes some errors. With careful reading almost everything can be fully understood.

C Distinctly limited range of expression with many errors. Given very careful reading, he can be almost fully understood within his limitations.

D Makes a considerable number of errors within a very limited range of expression, but demonstrates some ability to produce simple sentences so that with very careful reading most of what he writes can be understood.

E Has little or no ability to express himself in writing, even at the very simplest level.

The pieces range in quality from Grade A to Grade E pieces of

writing and each is commented upon. The assessor is expected to obtain similar picture sequences or devise equivalent writing tasks.

Candidates should next be given 'some passages' to read silently. They should then discuss the passages and answer questions put by the assessor. These answers may be spoken or written. Four sample passages of increasing difficultyof the narrative and expository

types are given. The assessor is expected to choose similar unseen passages to give to the candidates. No sample questions or answers are offered in guidance.

Listening comprehension should next be undertaken and tape-recordings of native-speaker conversations are suggested as being suitable for listening to. No sample conversation is given and consequently no sample questions or answers are given either. The assessor is, however, advised to establish whether a candidate has followed the gist of what he has heard and also whether he has heard the exact words of randomly chosen phrases. The assessor is again required to choose the recorded material to be used.

By this time the assessor may have a fair idea of the candidate's ability to speak English but he should encourage the candidate to discuss some further topics. The only topics suggested by way of example are connected with the proposed visit to Britain, and the assessor is advised to steer the discussion in such a way that a variety of different tenses have to be employed by the candidate. A tape is available with twelve short samples of candidates' speech. This is accompanied by a brief note on each sample and a suggested grade. The grades range from a very good A to an E.

3.3.4 There are no grounds for believing that these instructions are not followed in general terms by the majority of today's assessors. However, the responsibility for the assessment is entirely that of the assessor. He or she must choose or devise the material to be presented to the candidates, administer it and evaluate the performance in each of the four skill areas. The

validity of each assessment is dependent on the calibre of each assessor, who is normally either a British Council officer of junior or intermediate level, or a British teacher on contract with the Council, or a qualified person locally engaged by the Council (Hindmarsh, 1977: 24). Some assessors may test a number of candidates during a year and may have developed their own tests and techniques. Others may test very infrequently and use procedures which are not particularly well thought out, or simply use the suggested tests in the OSI. Moreover, applying the criteria in Table 3.5 may not always be easy. The descriptions are brief and not comprehensive. One assumption that seems to be made, in ability to speak English for example, is that as the accuracy, clarity, range of expression and fluency of the candidate decrease, so the ability of the listener to understand decreases and the effort he has to make while listening increases in proportion. This does not necessarily follow. It is frequently the case that a non-native speaker is perfectly intelligible, and that no effort is required on the part of the native speaker listener (who is resident in that country) in order to understand fully. This would appear to indicate an A according to the criteria in Table 3.5. But it may be that there are a number of pronunciation and structure inaccuracies which set the speaker some way from near-native performance and so indicate a B. The decision for the assessor is not always clear.

3.3.5 The reliability of assessments undertaken in this way is inevitably very low. With more than 5,000 candidates being assessed in one calendar year (the British Countil investigation reported

by Hindmarsh only gives numbers up to 1973, but indications are that numbers have continued to increase) several hundred assessors are involved. Consequently the chances of obtaining the opinion of the same assessor for more than one candidate in a given group of applicants for any one course in a British university are very low (Moller, 1977: 27).

No formal claims of validity for the BCSA are made. The assumption is that given the framework, the carefully worded definitions for the grade, and the experience and expertise of the many assessors, reasonably valid judgements will ensue. No follow up studies or applications of external criteria are known to have been made.

3.4 Consistency of BCSA with the Theory

It is hardly appropriate to talk of the theory behind these assessment procedures. They are typical of the type of test referred to by Spolsky as belonging to the pre-scientific period and display the characteristics listed in para 1.1.2. The only advance on traditional tests is that proficiency consists of displaying ability in the four main skill areas of listening, speaking, reading and writing. The instructions are specific on this point, and assessors are also instructed to give separate tasks to test performance in each of the skills. In traditional pre-scientific tests, assessors would have preferred to give just one overall grade to be arrived at as the result of an interview, during which listening, speaking and reading would have been assessed, and a short writing task. (Some hard pressed or less rigorous assessors still use that procedure but give three grades based on the interview!) Specific reference to the individual skills would not have been made,

and the main aim as quoted in 3.3.2 above would have been considered sufficient.

3.4.1 Further discussion on the BCSA's consistency with its underlying theory is thus unnecessary and irrelevant. There is no underlying theory stated but a strong appeal to traditional practice. The assessment is a set of pragmatic and expedient procedures which each assessor must carry out to the best of his or her ability in the light of the local situation. Certain basic principles have been laid down, and it is left to the assessor to adhere to them as best he can.

3.4.2 However, the extent to which the procedures are compatible with the current theory and practice is of more interest. It has already been noted that reference to specific skills has been made in the instructions. This is almost certainly due to the influence of the psychometric-structuralist trend, to use Spolsky's terms again, in which proficiency was broken down into its basic components and which put greater emphasis on testing the language skills. However, with the stress on 'performance' and 'participation' (see para 3.3.2 above), the procedures are not inconsistent with the current interest in communication displayed by linguists and language teachers. The definitions to be used for the grading (see Table 3.5) refer to communication. In understanding of spoken English, the candidate must display understanding when <u>being addressed</u> (the writer's underlining). In speaking English, <u>intelligibility</u> and the reaction of <u>the listener</u> are prominent, and in writing English, <u>intelligibility</u>

and the reaction of the reader are important. Throughout the procedures the assessor is urged to keep in mind the sociolinguistic environment in which the candidate hopes to use his English. Although no technical terminology is used in the instructions and suggestions, assessors are asked to assess the skills that testers of communicative competence aim to test. However, the approach is not consistent with current practice. There is no control over the material or content being tested, and the consequent unreliability of the procedures is still unacceptable. But if a concerted effort were made to develop and try out a comprehensive range of materials for presentation to the candidates, and if the performance definitions for grading were revised, these procedures could be made more consistent with current thinking on language testing than the theory on which EPTB is based.

3.5 Conclusions

In this chapter the design of EPTB and BCSA have been examined and the consistency of the theories underlying the two sets of procedures considered. In the light of the above discussion the following conclusions relating to the hypotheses stated in paras 2.9.1 and 2.9.2 may now be put forward:

(a) the construction of the English Proficiency Test Battery is consistent with the theory on which it is based, but is not basically compatible with current thinking in linguistics and language testing;

(b) the construction and administration of the British Council Subjective Assessment procedures cannot be said to be consistent with the theory on which they are based, since this is nowhere

stated. The procedures are, however, consistent with their general stated aim; the general approach which can be inferred from these procedures does contain some elements which are compatible with current thinking in linguistics and language testing.

CHAPTER 4 THE SAMPLE AND THE CRITERION MEASURES

4.0 Preparatory Phase of the Enquiry

Having established - or re-established in the case of EPTB - that both EPTB and BCSA exhibited consistency within their own frames of reference (although the absence of control of the content and the unreliability of BCSA rendered its consistency almost meaningless), the extent to which the two measures predicted the candidates' adequacy in English while pursuing their studies in Britain had to be established.

To achieve this a study was set up in which the performance of a group of students was observed firstly on the measures in their home countries and later on some special measures administered while they were studying in Britain. The preparatory phase of the study involved

(a) identifying an appropriate sample of students who were not native speakers of English and who had taken either EPTB or BCSA,

(b) identifying or devising criterion measures to which the sample were to be submitted.

4.1 The Sample: Considerations and Constraints

When the sample for the study was being selected, one theoretical problem had to be resolved - the size of the sample. Linked closely to this decision were two practical considerations - the availability of subjects and the availability of data. Determination of the sample, and to a lesser extent the nature of the criterion measures, was also influenced by the time and financial resources available. The study had to be limited to the academic years 1973/4 and 1974/5, and funds covered only essential travel and occasional secretarial assistance. But on the other hand the full cooperation of relevant departments of the Home Division

of the British Council in London was enjoyed, and this made the enquiry possible.

4.1.1 The first major consideration in selection of the sample was size. Since the aims of the two proficiency measures were the same, it was argued that the criterion measures would be valid for the whole sample regardless of the assessment procedure initially undertaken. But since there were two procedures under investigation, there would necessarily be two sub-samples, one having taken EPTB in their own countries and the other BCSA. In order for results of the enquiry to be significant, and in anticipation of inevitable wastage for various practical reasons, it was decided that each sub-sample should initially consist of at least 300 subjects, thus producing a minimum total sample of 600.

4.1.1.1 A further consideration was whether the sample should be a representative sample or a random sample of all candidates who had taken the assessment tests. When the enquiry was begun in the autumn of 1973, the writer learned that the British Council then administered over 3,000 award holders or private students from overseas in Britain, but that of that number only a third would have had English proficiency assessments. It was then decided that all cases would be included in the sample with a maximum for practical reasons of 1,000. A sample of between 600 and 1,000 would then constitute an acceptable, though not necessarily representative, sample of all those taking the tests in a single year. The total number of English proficiency assessments by the British Council world wide for 1972 was 7,259 (Hindmarsh, 1977).

Thus the sample proposed would represent between 8% and 14% of the total applicants assessed overseas in the preceding year.

4.1.1.2 Another factor affecting the sample was range of English proficiency. For obvious reasons the only subjects available in Britain were those that had been accepted by universities or other institutions. Presumably the college authorities or the British Council had deemed the English proficiency of all the subjects to be acceptable. Ideally the design of the enquiry called for subjects whose English proficiency was <u>inadequate</u>. But it was hardly likely that either the British Council or British universities would willingly accept students whose English was known to be inadequate for their studies! Further it was not possible for the writer to fund such students, and human and ethical considerations also militated against such action! But this consideration was not ignored. There were grounds for believing that for a variety of reasons, some students with inadequate English on their EPTB or BCSA assessments had managed to gain admittance to tertiary courses. It was therefore clear that this category of student, theoretically essential for the sampling, would be represented. In the event the number proved higher than expected!

4.1.2 The second consideration was that of availability. Since EPTB was restricted to use by the British Council, or for use by other groups under the supervision of the British Council, and since the subjective assessments were conducted solely by the Council, all the subjects and the relevant data would be available through the British Council. This consideration was a distinct advantage,

but applied only to students sponsored by the British government, the British Council or other bodies for whom the Council acted as agents eg certain UN agencies. A large number of candidates assessed overseas by the British Council continued either to Britain or to other English-speaking countries independently of the Council. It would have been a slow process tracing these people, and in view of the already adequately sized sample, these possible subjects were ignored.

The subjects for the study had to be available during their first few months in Britain for the follow up measures, and this aspect of availability was more complex. Subjects were liable to be scattered all over London and spread from Torquay to Aberdeen. Those who received funds through the Council were the easiest to contact and constituted well over 90% of the sample. However, it was necessary to obtain the cooperation of the regional offices of the British Council throughout Britain, and this cooperation was always forthcoming.

4.1.2.1 One of the constraints on availability, particularly with postgraduate students, was that field trips, study tours and attendance at conferences, as well as the regular timetable demands limited the opportunities for contact between the subjects and the investigator. This meant that contacts had to be carefully planned and the cooperation of tutors and other academic staff was essential.

4.1.2.2 Cooperation of the subjects in respect of the criterion measures was vital. The investigator had no authority to compel

any subject nor any tutor to participate. Cooperation had to be on a voluntary basis. Consequently it was decided that when cooperation was requested, the request should be accompanied by a full statement of the purposes of the research. For the most part the topic was one which created a lot of interest, particularly amongst academic staff and admissions tutors. When students arrived to participate in the special tests, a full explanation was given to them first, and cooperation was then readily offered.

4.1.3 Linked with availability was the question of length of stay in Britain. The most common period of study was nine months, one academic year. However, many students working for research degrees planned to stay for two or three years, while a growing number came for special courses of three to six months' duration. Many had also come for short attachments of less than three months. Most, however, were linked to the academic year. Since the purpose of the enquiry was to relate their performance on EPTB or BCSA to the adequacy or otherwise of their English for their specific training, it was necessary to allow time from the beginning of their course to elapse before any valid assessment could be attempted. It was felt that at least one term should be allowed. By that time students should have overcome initial difficulties arising from their new environment and the often different approach to study. By that time tutors too should have had the opportunity to come to know their students' abilities. Evidence that students themselves recognise language difficulties in their first months was obtained in a survey of students from overseas conducted in the mid 1960s. Their major perceived difficulty was in understanding. Only, 10%

of those surveyed indicated long periods of difficulty. The majority, therefore, felt they had settled down adequately (Morris, 1967: 24-25). The present study was concerned with measures to be administered once the initial settling period was over. Three months was judged to be the minimum period to be allowed.

There was a further practical consideration. Collection of the initial data for such a large sample, locating subjects and distributing the criterion instruments would take all of those three months, since the work was to be carried out by a single investigator. Time also had to be allowed for latecomers to arrive, and so the initial collection could not start until late October.

The presence of the subjects was also necessary for the administration of the criterion measures and this had therefore to take place during the second term. But more important was the fact that both the proficiency procedures being investigated had been devised primarily for students wishing to follow academic courses in Britain, and these students were most likely to stay for at least eight months. If the minimum period of stay were three months, the proportion of subjects following non-academic courses was likely to be substantial. To ensure that the majority of subjects was oriented towards academic courses, and to take account of some of the practical constraints of time, it was decided that the sample should consist of subjects studying in Britain for at least six months. It was also decided to administer the criterion measures as far as possible after the subjects had been following their courses for at least three months and before the end of their sixth month. This decision eliminated students staying in Britain

for very short periods and also eliminated those who had already completed one academic year or more.

4.1.4 A further consideration was whether to control for variables such as age, sex, country of origin, course of study. Since the initial decision was to obtain a sample approaching 1,000 it was not felt that controlling for these variables served any useful purpose at that stage. A large sample of students was almost certain to encompass a sufficiently wide range that sub-samples based on one or more of these variables could be created and examined at a later stage in the enquiry. It was considered essential, however, to have certain items of information on each subject before he or she was included in the enquiry. The background items considered essential were

- country of origin
- pre-departure language assessment (EPTB or BCSA)
- course and place of study in Britain.

Without this information the enquiry could not be carried out. Age and sex were also recorded. The highest level of education already attained was considered desirable as it might be related to the level of English prior to departure, and the level of course to be followed in Britain was thought to be related to the level of English required on the course. The number of weeks of remedial English undergone on arrival in Britain was also recorded.

4.1.5 After consideration of the above factors it was decided that the sample should include all students in Britain administered by the British Council whose proficiency in English had been assessed in their home country either by EPTB or BCSA and for which the

scores or grades were available, and who had embarked on a course of study or training lasting at least six months but who had not yet completed one academic year. The size of the sample was to be subject to a maximum of 1,000. As many of the desired details of background information as possible were to be collated.

4.2 The Criterion: considerations

Identifying suitable criterion measures involved determining what exactly they were to measure. Two major possibilities were considered - success in the course being followed, and the adequacy of English for completion of the course. It can be argued that the distinction between the two is fine, or even unnecessary. The prime concern of a student is to complete the course successfully, and this is also the main concern of the sponsoring agency and of the instructional staff. But the implications of this attitude are firstly that if academic success and a score on an English proficiency measure are seen to be closely related, the English test is considered a predictor of academic success, which was not the prime purpose of the test, and secondly, that proficiency in English is seen as a condition sufficient for academic success. There are, however, other factors relevant to academic success, as many studies have shown.

4.2.1 In the majority of previously reported studies of predictive validity of language proficiency tests, academic success has been the criterion. In paragraphs 2.5.3 to 2.5.3.9, thirteen separate predictive validity studies conducted in USA, Canada, Singapore and Britain relating to six different English proficiency tests were reported. In all but one of the studies academic success in

the form of grade point average, examination results or other indicators of academic achievement was adopted as the criterion.

The advantages of using academic success as criterion are that such information is normally readily available, particularly in North America, and is totally independent of the language test constructor and of performance on the language test. The GPA is a particularly convenient criterion, since it can cover as little as a semester's work or as much as four years' achievement. It uses a scale with small intervals, theoretically a 41 point scale from 0.0 to 4.0, thus lending itself to product moment correlations.

Using the criterion of academic success in Britain can pose a number of problems, however. There is no widely used criterion similar to the GPA in North America and other countries. The essential award at the end of a year's study is either a pass or fail. But in some institutions different classes of pass may be awarded eg first class, second or third. In others the occasional distinction may be awarded. There are many other courses in which passes are not awarded. Students merely complete the course satisfactorily, and criteria for satisfactory completion may vary considerably from course to course. Research students may be permitted to continue for a further year solely on the basis of the tutor's report. This lack of consistency in the nature of the final assessment of academic courses means that when the results of a large sample of students from various institutions are compared the only common variable is the pass/fail or completion/non-completion distinction. There is one further practical complication when using academic success as a criterion. Most results are declared at the very end

of the third term or even as late as the first week of the long vacation. This means that it is often very difficult for the external researcher to obtain details of end of first year results from a variety of academic administrations much before the new academic year.

4.2.2 No other studies so far undertaken are known to have used adequacy in English proficiency as a predictive criterion. It has the advantage of being the quality which the pre-departure measures have attempted to measure or predict. It has the apparent disadvantages of being difficult to define and not easily measurable.

It would appear that different disciplines may require differing degrees of adequacy in English on the part of non-native speaking students. It is likely, too, that different types or levels of course may require the exercise of different language skills to differing degrees. A Masters degree that is obtained as a result of course work may well demand a greater proficiency in listening and reading and in writing under pressure than a Masters course that is geared mainly to experimentation and research. Moreover, a degree course in the same subject but conducted in different institutions may require varying degrees of proficiency depending on the expectations of and standards set by the staff involved. A definition of general adequacy in English proficiency for academic study would therefore seem impossible to establish. It cannot be described in absolute linguistic terms. What does seem stable, however, is the notion of adequacy, and since the descriptions

will vary for different situations it will be clear to both staff and students in each situation whether a student's English is adequate or not and the extent to which his/her English is adequate or inadequate.

4.2.2.1 If adequacy in English for academic study cannot be defined then it will not be possible to measure such adequacy. If, however, it is accepted that the notion of adequacy in English proficiency exists for every situation then it will at least be possible to establish whether that notion of adequacy can be attributed to each student in each separate situation. There are a variety of ways in which the presence or absence of that quality can be established.

4.2.3 It was decided that this study should take as the criterion the adequacy of the subjects' English for completion of the course being followed for the reason that this was the prediction made by the proficiency measures being investigated. By accepting the notion of adequacy without attempting to define it precisely in linguistic or behavioural terms it then became possible to develop some instruments to obtain this information.

4.3 Sources of Information Considered

In order to get reliable and valid information on a student's adequacy in English more than one source of information needs to be explored. There are three main sources in the type of situation being investigated. Firstly, there is the student himself, secondly his tutors, and thirdly an external assessor. Suitable instruments for these are a self-assessment rating, rating by the tutors and some form of

language test battery. Each of these possibilities was considered.

4.3.1 The principal question to be resolved when contemplating some form of student self assessment is to what extent can the information be relied upon? In other words, will the student be too lenient or will he be too critical of his abilities in English? Experience has tended to show that weaker students over-estimate their abilities whereas the proficient ones are still aware of deficiencies and tend to underestimate their abilities. Jordan (1977: 14) presents striking evidence to support the situation where students overestimate their abilities. In 1972/3 and 1974/5 in the University of Manchester students' self-assessment in English ratings on entry to their courses were compared with their scores on English proficiency tests devised by Chaplen (1970).

> 'Overwhelmingly the results showed that students at the lower end of the scale in the tests grossly over-estimated their language ability, especially in writing. ....of 144 students... 34 students obtained a score of between 10% and 40% (ie sufficiently low to warrant full-time English tuition for several months). However, 50% of these students described their written English as very good (6%) or good (44%); 41% described it as weak and only 9% as bad - the most apt description.'

It is also possible that it is in a student's interests to be optimistic, or pessimistic, about his abilities. For example, if a student suspects that poor performance in English may jeopardise his acceptance for a course or for continuing for a further period, he will tend to be optimistic. If on the other hand there is the chance of an extra month's free tuition in the country in which he is studying, he may well wish to underestimate his abilities in order to secure some extra tuition which he may not really need.

4.3.2 Tutors' ratings have the advantage that the tutor is not so involved personally with each individual's ability in English, and provided that he does not have too large a number of students to deal with he is likely to judge whether a particular student's English is adequate for the course or not. Although the tutor will in most cases be a specialist in a subject other than English, he will have a reasonably accurate idea as to the adequacy or otherwise of his students' competence in English. A tutor of English, on the other hand, will only know how his student is faring in the language classes and can only speculate as to the student's performance in the course of his other academic work.

Tutors' ratings give rise to some difficulties because there will be more than one tutor concerned with teaching each student, and opinions may differ. A further factor is that it may take a term or more for a tutor to be aware of individual abilities in English. However, an aggregate or average of different tutors' ratings can meet the first difficulty, and it is advisable to wait until the second term before rating students' English in any case - see the discussion in para 4.1.3 above.

Since neither students nor tutors are specialists in the assessment of language ability, the instrument used has to be devised in such a way as to enable the non-specialist to understand it easily and to respond meaningfully. The most satisfactory instrument is likely to be a questionnaire in which the responses to the questions are controlled and presented to the respondent with a series of alternatives.

4.3.3 A set of language tests provides an external measure to confirm or modify the ratings made by the students and the tutors. It provides a common measure of linguistic performance for all members of the sample. This contrasts with the adequacy ratings where individual 'adequate' linguistic performance will vary from institution to institution and even from individual to individual. The use of such a set of language tests means that actual linguistic performance of students with the same adequacy ratings can be compared, and the range of actual linguistic performance within any one segment on the adequacy rating scale can be observed. The degree of overlap in actual performance between different points on the adequacy scale can also be noted.

There are however two serious problems with administering language tests. The first is practical in that such tests are bound to consume a lot of time - particularly for the subjects who have to be persuaded to give up at least one to two hours of their own time to be subjected to them. In addition there are logistical problems of administering the same test to up to 1,000 subjects within a very short period of time at probably over a hundred different places in the country. The second problem is a theoretical one for the investigator. If the test is to be truly external, a decision has to be made as to what constitutes 'adequate' performance, and what kind of performance is to be deemed 'inadequate'. This decision as to where to place the 'cut off' or 'cut offs' must be a carefully argued but essentially subjective one and has to apply to all subjects regardless of their field of study, type and level of course, or institution. Subjects will thus be measured against a notional or 'typical' non-native English-speaking student.

4.3.4 In deciding what instruments to use for the study it was felt that one external instrument was essential, as was at least one internal instrument. Although theoretically it is more satisfactory to employ both the self-assessments and the tutors' assessments, it was considered that in view of the limited resources available one would suffice to give the desired information - and that the tutors' rating was to be preferred on the grounds of greater potential reliability. As has already been indicated above, self-assessments tend to move towards the central point of the scale and individual students will on the whole be less practised in assessing their own and others' English ability. Tutors are liable to be more experienced in making assessments, however informal, and reliability checks can be built into the questionnaire.

4.3.4.1 The choice of an external test presented another problem. Should an existing test be used, or should a new set of measures be devised? Such a test should clearly be based on a theory of proficiency consistent with current thinking and attitudes. A glance at the characteristics of existing standardised and 'validated' tests (see Table 1.1 above) shows that few tests have productive performance components and that most are based on theories current in the early 1960s in the same way as the EPTB. The Chaplen Test (1970), not included in Table 1.1 but available, was thought to be inadequate as it consisted only of multiple choice grammar and vocabulary tests and had no productive components. The conclusion was that a new set of language measures was to be devised.

4.3.5 The instruments chosen to determine the adequacy of the

sample's English proficiency were therefore

(a) a tutor's rating questionnaire and

(b) a short battery of English tests.

No existing instruments were known to be satisfactory for the purpose of the enquiry and so both had to be devised by the investigator. Tutors' ratings were to be sought for every member of the sample, but since the English language tests were to include productive components, testing each member of the whole sample would present severe practical and logistic problems. It was felt that it would not be possible to administer the tests to more than about 100 students. This would, however, constitute just over 10% of the year's sample and provide sufficient information for the purposes of the study. A decision was therefore made to test a sub-sample of at least 10% of the second investigation sample.

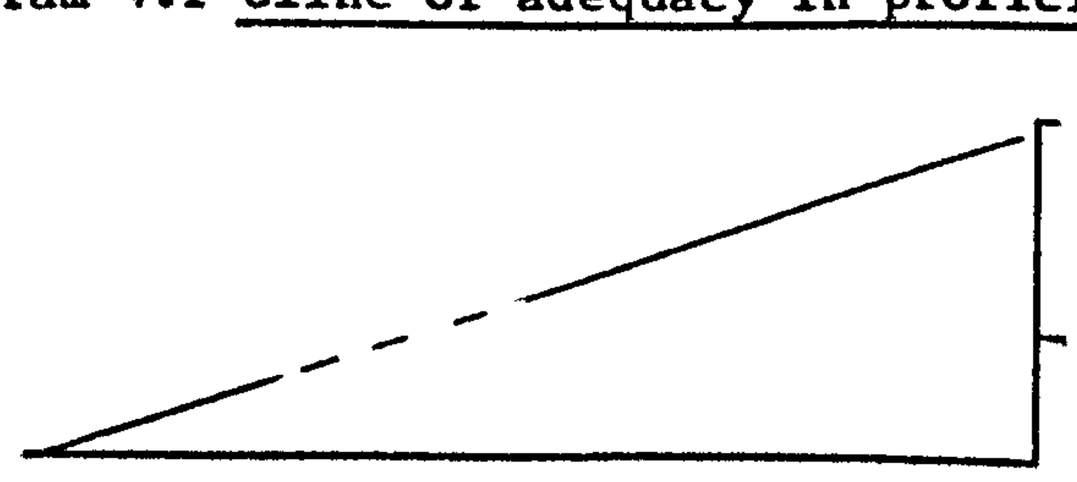
4.4 English Ability Rating Questionnaire

It has already been argued (paras 4.2.2 to 4.2.3 above) that it is possible to establish whether students' abilities in English can be considered adequate or inadequate as a result of ratings by tutors with knowledge of each individual student. An English ability rating questionnaire for completion by tutors was therefore devised. Three main principles governed the preparation of the questionnaire. Firstly the information to be obtained had to be decided and made specific. Secondly features had to be built into the questionnaire in order to maximise the reliability of the responses. And thirdly, the questionnaire had to be simple and brief so that it could be completed with ease and speed, thus making it as light a task as possible for respondents and so facilitating a high level of response.

4.4.1 Tutors were required to fill in one questionnaire for each student. The basic question to be put was 'Is the English proficiency of this student adequate for undertaking specialised studies or research in his field of study?' The simplest possible answers to this question are 'Yes' and 'No', but it was felt that these were over-simple. There would be a majority of cases where it would be quite clear which would apply, but in other cases it would be very difficult to decide if the student's English was in fact adequate. Moreover the difference in abilities between someone who was judged, on balance, to be just adequate and the person judged adequate with no hint of doubt in the mind of the assessor would be totally hidden.

4.4.1.1 There is no very clear demarcation or cut off between inadequate and adequate proficiency in English, or indeed in any language. In absolute terms adequacy can be represented as a cline from total inadequacy to complete adequacy. The crucial area of the cline for the purposes of the present study was considered to be the 'borderline' area where doubts about the presence of adequacy are known to exist. A learner does not have to develop language

Diagram 4.1 <u>Cline of adequacy in proficiency in English</u>



proficiency very fully before reaching a stage at which his proficiency is almost adequate for certain tasks. Further

development in proficiency will enable him to operate with minimum adequacy, but he will with difficulty be able to achieve tasks that he has set out to achieve. The learner's syntactic, phonological and lexical competence is extremely limited, but a threshold will have been reached from which he will be able to advance with greater and greater communicative effect. This is exemplified in the work undertaken by the Council of Europe in the development of a Threshold Level to be attained by learners of European languages (van Ek, 1975). The majority of the learner's progress towards the goal of native-speaker-like proficiency is spent increasing the extent of his competence and the adequacy of his performance.

It is for these reasons that the area of uncertainty, or stumbling adequacy, is not right in the middle of the cline but somewhere between a quarter and a third of the way along from the bottom. Over half the cline represents varying degrees of adequacy. It was therefore felt that this area could be explored and that respondents should be given the opportunity to rate their students highly where appropriate. It was also felt that the designation inadequate would also be insufficient, and that it would be advisable to give respondents the possibility to differentiate between those students who were very close to the borderline area, but still not quite there, and those whose proficiency was clearly a long way from being termed even just adequate.

4.4.1.2 It was decided that respondents should reply to the basic question twice - firstly by choosing one of three responses relating to the three areas indicated on the cline in the diagram, and

secondly by choosing one of a larger number of responses which would seek to establish a little more precisely the extent of students' adequacy or inadequacy. The alternative responses to question 1 were

A. - more than adequate

- adequate

- not adequate, and

B. - Shows native speaker ability

- Clearly a non-native speaker because of minor faults in English usage, but this does not handicap him in his studies

- Makes many mistakes in English usage, but this constitutes only a minor handicap for him in his studies

- Shows many weaknesses in English usage but his English ability can be considered just adequate for his studies. A higher standard is desirable.

- Shows considerable deficiencies in English usage, which constitutes a handicap for him in his studies. A higher standard is necessary.

- Shows very little ability in English and is well below a satisfactory standard.

The first three responses at A correspond directly with the three sections of the cline, as in Diagram 4.1.

The six responses at 1B represent an expansion of the responses at 1A. Reference is made either to general ability in English (eg little, native-speaker), or to the presence or absence of one or two features of performance, or to the need for improvement, or to a combination of these.

The first of the performance features is inaccurate use of English. In the responses the words <u>faults</u>, <u>mistakes</u>, <u>weaknesses</u> and then <u>deficiencies</u> are used in this context. They could be said to represent points on a cline from no inaccuracies to total deficiency.

The second feature is the handicap that the persistence of these inaccuracies may constitute for a student in pursuit of his/her studies. There is a nearly parallel cline of responses from no handicap to complete handicap.

The need for improvement in English is stated or inferred in the responses. In some cases, improvement in English is not necessary. In others it is desirable, while in others it is essential.

4.4.1.3 Table 4.2 below sets out an analysis of the contents of the responses to question 1B. For the purposes of the analysis values are given to each of the responses, with 6 indicating the highest ability description (shows native speaker ability) and 1 indicating the description of the student with little or no ability in English. At the highest level of ability (6) no reference is made to inaccuracies, handicap or improvement. Clearly these characteristics do not apply at that level. But the characteristic of inaccurate use appears first at level 5, the identification of some kind of resultant handicap at level 4, and the need for improvement only starts to be relevant at level 3.

Table 4.2 Analysis of contents of the responses to question 1B of the English Ability Rating Questionnaire

Level	Inaccuracies	Handicap	Improvement
6	nil	nil	nil
5	minor	nil	nil
4	many	minor	nil
3	many	major	desirable
	Deficiencies		
2	many	definite	necessary
1	complete	complete	necessary

The responses in question B were not meant to be directly linked to the three responses in question A. However, it is clear that the last two (levels 2 and 1) relate to the response - <u>not adequate</u>, that level 3 certainly corresponds to the response - <u>adequate</u>, and that levels 4 to 6 correspond to the response - <u>more than adequate</u>. The response indicating that the individual's mistakes constitute a minor handicap (level 4) could be construed by some as relating to the just adequate category, particularly as in the layout of the form the two responses are adjacent, and the impression was given that two responses in question B corresponded to each single response in question A. (See Section 1, Appendix I). This weakness in the format was rectified for the second investigation.

4.4.1.4 The second set of responses at 1B served another important purpose - that of checking the reliability of the responses at 1A, for in practice respondents were answering the same question twice and within a matter of seconds. The expectation was that responses to both A and B should correlate highly.

4.4.2 The second question to be asked was considered to be a more detailed version of the first. Instead of being asked to rate each student's <u>general</u> ability in English, respondents were next requested to assess four specific abilities in English:

the ability to understand spoken English
the ability to speak English
the ability to understand written English
the ability to write English.

Since the second question (No 2 on the questionnaire) was really

the same question as No 1, but with more detail requested, it was felt that another, objective, format should be used. The difficulty with using statements or definitions as in question 1 was that since the statements could only be general, they would almost certainly have to be expressed in the same terms as those in question 1. A further purpose for this question was to give respondents the opportunity to differentiate between subjects' abilities in different language skills, eg to show whether they judged the listening ability to be the same as or different from the speaking ability of a particular student. It was further intended to give respondents greater freedom of response in this question than they had had in the first. This need also argued against the use of alternative responses as in question 1.

4.4.2.1 The most suitable format, therefore, seemed to be one in which the respondents were to indicate their assessments without recourse to the use of words but within defined limits. Consequently it was decided that each of the four linguistic abilities should be represented on the questionnaire by straight lines, or clines, and that the respondents should indicate with an X where along the ability cline they judged the student's ability to lie. The cline was divided into three sections with indicators marking the centre of each section. The sections were marked <u>not adequate</u>, <u>adequate</u>, and <u>more than adequate</u>. The scale was not meant to be precise but the labels offered guidance. The use of this format made it possible for respondents to indicate the general adequacy of a student's English in question 1, to differentiate between competence in different skills, and to indicate a student's

inadequacy, or near adequacy, in a particular language skill when applicable. Question 2 is set out in Table 4.3.

Table 4.3 Question 2 of the English Ability Rating Form

2. Please consider this student's ability in the following language skills and where possible give your rating of his ability in each skill by putting X at an appropriate point along the scale from zero ability to native speaker ability.

	not adequate	adequate	more than adequate
Ability to understand spoken English			
Ability to speak English			
Ability to understand written English			
Ability to write English			

4.4.2.2 There remained the question of how to interpret the responses, since no points were clearly defined. However, each section was divided into two further sections and it was decided to award one point per section. Thus 1 and 2 were awarded for not adequate, 3 and 4 for adequate, and 5 and 6 for more than adequate. The scores for each skill could then be summed up and an overall score for all abilities obtained. Although this total is not meaningful in itself it would enable a comparison to be made with the general assessment made in question 1. For example, if a student had been rated adequate in question 1A or level 3 in question 1B, it would be reasonable to expect the total of the ratings in question 2 to be between 12 and 16 inclusive. Any score

significantly outside this range might indicate a certain inconsistency in the rating or, more probably, that the rater had given greater weight to one or two of the language skills, which would be reflected in the scoring on the clines.

The assessment of the individual language skills served to confirm or modify the general assessment made in the first question. It also gave information of general skill areas that tutors identified as being either most or least adequate.

4.4.3 One important factor in the study was the amount of improvement in English experienced by the subjects. It was only to be expected that after 3 to 6 months in Britain, students would have made some evident improvement in their English. It was assumed that tutors would be aware of the absence or presence of this factor. Therefore respondents were asked to indicate what kind of improvement in English ability they thought the students had made since the previous October. Three alternatives were given:

- considerable improvement
- a little improvement
- no improvement

4.4.4 At this point the questionnaire was trialled. As the first academic term was still in progress it was thought that a proper trial with tutors from different departments might not be appropriate as they would not yet know the ability of their students' English and would have had few opportunities to assess it. However, some overseas students in the University of Edinburgh were following in-session remedial English classes as a result of

having been referred to the Department of Linguistics by their departmental tutors. The range of ability on entry to the course was quite wide as measured by ELBA at the beginning of the term. Scores ranged from 36% to 68%.

Their English tutors, postgraduate students in the Department of Linguistics, were asked to complete the questionnaire on the assumption that they were also academic subject tutors. It was explained to the three tutors that the purpose was not necessarily to obtain accurate information but to test the design of the questionnaire. Fifteen students were rated by the three tutors. In many cases it was not possible to rate writing ability since insufficient writing had been done. In almost every case differentiation was made between different skills and there was generally close correspondence between responses to the two sections of question 1.

4.4.4.1 As a result of the trial changes were made in the questionnaire. In question 1A it was decided to revise the labels for two of the boxes. 'More than adequate' was found to be unsatisfactory and not really a likely category. Completely adequate, indicating mastery sufficient for all demands was considered more appropriate. The second category adequate was thought not quite satisfactory. It was to cover the category of student whose English was either only just adequate or a little more than that. It was therefore changed to just adequate. It was also decided to recognise the existence of female students by changing 'him' and 'his' to 'him/her' and 'his/her'.

There was a major change in question 2. The organisation of the clines in sections parallelled question 1 too closely. It was decided therefore to exclude all divisions and to label only the end points of the clines and one intermediate point as a rough guide. In addition the tutors recommended that the top of the scale - completely adequate - should be at the left, next to the captions. These changes were incorporated. Extremities were labelled completely adequate and inadequate, and just to the right of the mid-point the label adequate. In this way the relationship between questions 1 and 2 was much less obvious and raters were given the freedom to place anywhere on each cline. The only limits were the end of the cline, which contrasted with the fixed alternatives in question 1.

The change in question 2 entailed a change in scoring procedure. The break up into 6 sections was felt to be still too crude and too close to question 1. It was therefore decided to extend the length of the clines from 10cm to 12cm and to attribute 1 point to each centimetre of space. Thus each cline acted as a 12 point scale. When the questionnaires were completed, it would be possible to take a centimetre rule and attribute a point on the scale for each of the skills assessed, eg an X marked between 6 and 7cm from the right of the cline would be awarded 7. Comparisons with question 1 would be effected. A student rated as just adequate and level 3 in question 1 might be expected to obtain total ratings of between 17 and 24, the total ratings possible now being 48.

4.4.5 Two further questions were added to the questionnaire. Allied

to the question of improvement was the question of further tuition specifically in English. It was thought that this might have some bearing on a student's improvement (question 3) and possibly on his adequacy. Consequently tutors were asked to indicate in question 4 whether the student had received any tuition in English since the previous October, with three possible responses:

- Yes
- No
- Don't know.

This particular information was not always readily available to the researcher or to the British Council. In many cases not even a tutor was aware of whether his student had had this tuition or not.

4.4.6 The fifth question required the tutor to write down the degree, diploma, or other qualification the student was working towards. Again this information was not always readily available to an outsider, as occasionally students changed courses at the last moment, or even in mid-term. The sixth amd final question was an invitation to the respondent to make any further comments on the reverse of the form.

4.4.7 Each English Ability Rating form had four identifying pieces of information, the student's name, country, university and field of study. This information enabled the British Council to identify the appropriate tutor easily and enabled the tutor to complete the form without having to be concerned with administrative details.

4.4.8 The revision of the form and the addition of questions four, five and six meant that the questionnaire could no longer be

accommodated on one side of paper. It was therefore decided to keep the instructions as basic as possible and to provide a key on a separate sheet of paper (one side only) containing the full instructions and further guidance as to how to complete each question. Thus the new instruction for question 1 became

> 1. General ability in English
> (Put X in the appropriate box in Column A and in column B)

The accompanying key amplified this:

> 1. Is the general ability in English of the student (ie present ability, not potential) adequate for undertaking specialised studies or research in his/her field of study?
>
> Please answer this question by putting X first in the appropriate box in column A and then in the appropriate box in column B.

4.4.9 It was felt that the form at this stage satisfied the criteria laid down in para 4.4 above. The information requested related first to general English ability (question 1) and more specific abilities (question 2). Further relevant information for the study was also requested, as it was felt that the tutor - possibly after consultation with the student concerned - was in the best position to provide it. Finally space was given for further comments which could lead to accounting for factors not covered, and which might lead to revision of the instrument for the follow-up. Reliability had been built in using the test/re-test principle. In question 1 the same question was asked twice, and the total score in question 2 could also act as a further indicator. It was felt that the format was readily intelligble, as the few tutors who used it on a trial basis had found it satisfactory. It was brief and simple in that all the responses were set out on one side of paper only. A letter to tutors explaining the scope of the enquiry and requesting

their help was also drafted. It too was limited to one side of paper only. The trial rating form, the revised form, the key and the letter to tutors are to be found in Appendix I, Section 1.

4.5 Communicative Proficiency Measure

The second instrument to be devised for the investigation was a short battery of English tests, referred to hereafter as the Communicative Proficiency Measure.

The purpose of the Communicative Proficiency Measure (CPM) was to measure each student's ability to communicate and to understand communication directed to him or her. Communication was intended to relate to 'real life' situations more than had been the case with previous language proficiency batteries, and the assessment of linguistic ability eg knowledge of the grammar, range of lexis, accuracy of pronunciation, was to be secondary.

4.5.1 With the emphasis on communication, it seemed appropriate to develop integrative tests in which the student had to interact with a listener, a reader, or simply with the text, rather than to develop discrete items or single skill linguistic tests. Nevertheless one language skill or mode would predominate in each subtest. Another factor to be considered was the kind of proficiency to be measured. Four types of proficiency test were discussed in paras 1.4.1 to 1.4.4 - examiner based tests, language based tests, proficiency based tests and function based tests. Examiner based tests were clearly not suitable for the present study. It was argued that language based tests and proficiency based tests

measured overall proficiency and that function based tests measured functional language proficiency. In the present context of determining the adequacy of students' English for their studies, tests of functional language proficiency seemed most appropriate while tests of overall proficiency offered another dimension to the language performance being measured. The proficiency based tests and functional based tests were seen to have further relevance to this study as their format was integrative and skill based.

The major constraint to take into account was the factor of time. It had already been decided to administer the language tests to only 10% of the total 1974/75 sample (para 4.3.5 above) or approximately 100 subjects. These subjects were located in different parts of Britain and would be giving of their own time to make their way to and attend the testing sessions. Moreover only the investigator was available to undertake the administration and marking of the measures. Everything pointed to making the measures as efficient as possible so that they could be completed within a period of one and a half hours, but certainly not more than two. The tests had also to be susceptible to rapid and, where appropriate, objective scoring.

4.5.2 The choice of subtests was governed by four requirements - the need to assess candidates' ability to commmunicate with other speakers and writers of English, the suitability of tests of overall and functional language proficiency, the need to sample a maximum number of items in as short a time as possible, and the need to find common tasks suited to the candidates' range of academic interests

and language ability. A number of options were considered and discussed below in terms of type of proficiency tested, language skills tested, test type, number of items or range of sampling, time required and type of assessment.

4.5.2.1 Oral communication can be assessed by means of a test of functional proficiency involving the speaking and listening skills (predominant skill underlined). The normal test type is an oral interview. In the context of the present study there were no set number of items, but the language sampling had to be extensive in terms of topic and opportunity for complexity of expression. The test lasted only about 15 minutes and was assessed subjectively.

The understanding of oral communication in which the subject takes no direct part but merely listens and responds or reacts can be tested in two ways. The first is functional, but because the subject does not participate, it is really a test of overall proficiency in a functional context. Only the listening skill, with a little reading, is involved. The test type is comprehension of spoken monologue, dialogues and discussions. 25 items would provide a sample of a range of semantic, syntactic, attitudinal and stylistic features in up to 20 minutes which could be objectively scored.

The second way is by a test of overall proficiency involving the listening and writing skills and using dictation. Two texts could be sampled yielding up to 120 items in about 20 minutes. The scoring would be objective.

4.5.2.2 Dictation provides the link with the assessment of written communication. A test of functional writing proficiency, involving the skill of writing only can be constructed using the writing task or essay format. The sample of language produced will not be very great, but adequate for the purpose, and the test should last about 30 minutes and be scored subjectively.

Assessing the understanding of written communication is not so easy. However a test of overall proficiency in which the candidate reads and interacts with the text, before writing individual words is available in the form of a cloze test. Four passages with 30 blanks each would provide a good sampling of texts and 120 items such that could be completed in less than one hour. The scoring would be objective.

4.5.2.3 Table 4.4 summarises the options and their characteristics.

Table 4.4 Characteristics of communicative proficiency options

		Proficiency type	Language skill	Test type	Number of items	Time	Assessment
Oral Comm	1	Functional	Speaking	Interview	-	15 min	Subjective
	2	Functional/ Overall	Listening	L comprehension	25	20 min	Objective
	3	Overall	Listening	Dictation	120	20 min	Objective
Written Comm	4	Functional	Writing	Essay	-	30 min	Subjective
	5	Overall	Reading	Cloze	120	60 min	Objective

Certain imbalances in this particular combination of subtests became apparent. Firstly, listening is tested twice. It has been argued that the type of listening comprehension test envisaged would be a test of both functional and overall proficiency which appeared

desirable. Yet it might only yield 25 items in about 20 minutes. This would indicate a relatively inefficient test and because it would be short, it might not be so reliable, nor might the sampling be very wide. Lengthening the test would take up more time. It has the advantage of being objectively scoreable, as does the dictation.

The second imbalance is the length of the cloze test both in time and in items. It was not known exactly how long it might take as it was initially felt that all candidates should be given adequate time to complete the test and so timing would be geared to the slower and less able subjects. It was felt that in order to sample a sufficient variety of text types a large number of items was necessary. However, it was clear that one hour could not be spent on the cloze test and that either the number of texts would have to be reduced or the number of items in them reduced - possibly to 25 each.

The shortest test was apparently the interview, although it was recognised that 15 minutes might represent a minimum if slower speakers were to cover the same range of topics as the faster speakers. Similarly it was felt that 30 minutes might be a minimum for the essay as it would still only represent a relatively small sample of written language from each subject.

The total duration of the measure was calculated at approximately 145 minutes, 2 hrs 25 mins. This was well outside the 1½ to 2 hr limits that had been imposed (see 4.5.1 above) and some changes in the plan had to be made.

The first decision was to drop the listening comprehension test. There were several reasons for this. The time taken to administer such a test was long in relation to the items involved. Similarly it would also take a disproportionate amount of time to devise and record, using a number of voices. But more important, it was not clear that it was necessarily a test of communicative proficiency to the extent that others of the proposed Battery were, eg essay, dictation. As it was also one of two testing through the listening mode it could be dispensed with as the subjects would still be required to listen to spoken English in both the dictation and the interview.

No other measure was dropped for the trial testing, but it was decided to reduce the number of items in the cloze test and to experiment with the timing during the trialling. The specifications and contents of the subtests for the CPM are discussed in the following paragraphs.

4.5.3 The interview.

The purpose of the interview was to provide opportunities for the subject to speak freely and at some length on a range of topics. The role of the interviewer was to 'provoke' the subject to speak and at the same time conduct some meaningful communication, so that at the end of 15 to 20 minutes, the interviewer would have observed a sufficient sample of the subject's speech in order to form a judgement about the quality of the performance.

To achieve this some basic structure to the interview was required, particularly as the subjects to be assessed came from a wide range

of family, national and academic backgrounds. It was also assumed that the sample encompassed a wide range of proficiency in English.

A feature of most interviews is that the interviewer, or interlocutor, is generally the dominant party in the discussion who directs the course of the exchange. However, since the purpose of this interview was to encourage the subject to produce as much free speech as possible the general tactic was for the interviewer to hand over as much control of the exchange as the subject was willing and able to assume as the interview progressed.

4.5.3.1 It was planned to achieve this by conducting the interview in four phases. The first phase would be purely introductory and largely monopolised by the interviewer who would explain briefly to the subject the purpose of the whole interview. Its secondary aim was to help put the interviewee at ease, a necessary procedure when the interviewer is not well known or a complete stranger to the subject, as was the case in this study. The main explanatory points to be made at this stage were to indicate the research context in which the testing was taking place, to explain that any judgements made would be kept confidential and sent neither to the institution where the subject was studying, nor to the British Council or other sponsoring body. It would also be important to stress that the 'results' would have no bearing whatsoever on the outcome of the subject's course of study in Britain. Thanks should also be expressed at this stage for the subject's cooperation.

4.5.3.2 The second phase of the interview resembled the classical

interview in that it consisted largely of 'wh' questions addressed to the subject. It was important at this stage that the subject should experience no difficulty in the content of his answers and so be able to concentrate all his attention on the formulation of his utterances. Also by talking about facts and people that were very familiar he should gradually have become more relaxed.

Accordingly, questions in this phase of the interview were directed to the following topics:

home country	- family
	- residence
	- education
	- work
	- English learning background
travel to Britain	- route and transport
	- reception by British Council
	- English tuition, if any
living in Britain	- kind of accommodation
	- meals
	- house rules, if any
	- communications with home country

opinions on experience to date, relating to any of above.

This was by no means an exhaustive list, and the interviewer had to be ready to pursue some topics further than anticipated, because the subject seemed keen to talk more about them. He had also to be ready to pursue topics which had not been anticipated but which would clearly produce a useful sample of speech. It will be noticed that the final group of topics relates to opinions which would be requested by such questions as "What do you think about...? How did you like ...? What is/was your impression of ...?" These questions, if answered properly, led to the use of more discursive discourse and more complex structures. They also provided a useful link with the third phase of the interview.

4.5.3.3 The third phase was that part of the interview which was designed to give the subject the most freedom of expression and which was linked to the 'real life' situation of having to explain work or ideas orally to tutors or to fellow students. He was asked by the interviewer to speak about important aspects of his research or course as follows:

- an account of the design of a research project, or
- an account or evaluation of a component of a course, or
- any other specialist topic or experience, and
- an estimation of the value of the work done once he has returned to his home country.

In this way every subject had the chance to give an account, make judgements, make prognostications and generally express a similar range of functions, although in almost every case the subject area was likely to be different.

During this monologue on the part of the subject the interviewer had to keep a check on the variety of the discourse and make attempts to change its direction if it was becoming too repetitive or if the style was too restricted. This was achieved by interjecting questions or comments at appropriate moments in an attempt to extend the range of language used by the interviewee. Even if the subject was producing a satisfactory flow of speech, the interviewer had to bear in mind that the main purpose was to assess communication and that in oral communication one of the speakers is rarely silent for long stretches of time, and that one very important feature of such communication is unpredictability, discussed by Morrow (1977). In order to introduce unpredictable (from the subject's viewpoint) elements into the exchange the interviewer should capitalise on his ignorance of the subject under

discussion, or feign ignorance, by unexpectedly asking for clarification of certain statements, explanations or definitions of particular technical terms, substantiation of certain statements, or for reasons why the subject disagrees with a view expressed by the interviewer. In this way the latter was able to judge to what extent the speaker was or was not 'thrown off linguistic balance' by these interruptions and to account for this aspect of performance in the final assessment.

4.5.3.4 The fourth and final phase was the passing of complete control to the interviewee by asking him to put any questions he might have to the interviewer. This phase had a number of objectives. At its simplest it enabled the assessor to note whether or not the subject had difficulty in forming questions in English. It also gave the subject the opportunity to question the interviewer further on an aspect of the current exercise which might have been worrying him, and it also gave him the opportunity to introduce a topic about which for personal or other reasons he was anxious to have information or somebody's opinion. In any event it was likely to extend the range of the sample of the subject's speech. If asked questions, the interviewer had to respond as fully as he thought appropriate, because that was the only part of the exchange during which the interviewee was encouraged to <u>listen</u> to a reasonable flow of speech. The interviewer was able to note whether the subject had any strategies for interrupting the discourse and the extent to which he had understood. If the subject had no questions, or only some very simple factual questions, the interviewer could abandon this phase of the interview or start

talking in some detail about a new topic in an attempt to form some further opinion of the subject's ability to understand.

The interviewer then had to bring the session to a close by making an appropriate gesture, eg putting away his pen or papers, standing up, referring to the time and his next appointment or to another waiting student. He had to make sure that he thanked the subject he had just interviewed.

4.5.3.5 One of the major problems with interview tests is the criteria by which performance is assessed. The interview, as outlined in the previous paragraphs, did not lend itself to a discrete item approach. Subjects were not being tested for the content of their responses, nor for the accuracy or precision of their language. These were just two factors in the overall communication that was generated, observed and assessed. The traditional method of assessing interviews is to give an overall impressionistic mark or grade but this method has been shown to be far from satisfactory - see the discussion of BCSA above (para 3.3.3).

4.5.3.6 One possible approach is to adopt an analytic method whereby certain features of the performance are isolated for assessment purposes and scored individually. Three problems arise with this method. Firstly there has to be a decision as to what features to test and evaluate, bearing in mind that if the list becomes too long it may prove very difficult for the assessor, when he is also the interviewer, to make valid judgements. Secondly the delicacy of the scale used for each feature has to be decided, taking into

account what it is feasible to observe in the time available. Thus a nine or ten point scale may be far too fine a scale when an assessor is to make up to ten different judgements in the space of approximately 15 minutes, and anything less than a five-point scale may be too crude to yield meaningful information. Thirdly there is the question of weighting the individual subscores and establishing what a total of these subscores might mean anyway.

In the context of the communicative proficiency interview the following were suggested as features which could be assessed individually:

i. extent of interviewee's comprehension
ii. extent of interlocutor's comprehension
iii. whether communication was effective
iv. fluency of the subject
v. appropriateness of the lexis used
vi. accuracy of syntax
vii. accuracy of phonological features

Features ii. and iv. to vii. represented assessment of the productive aspects of the subject's communication; i. assesses his comprehension of what is being said to him, and iii. is an assessment of the overall effect of the exchange.

4.5.3.7 An alternative approach is to adopt an overall performance scale with brief definitions of the type of performance represented by each point on the scale. Options are either a scale with an even number of points or a scale with an odd number of points. Thus a scale of 1 to 4 would represent two lower categories and two upper, while a scale of 1 to 6 could contain two points representing poor performance, two points representing average or medium quality performance and two points representing good

performance. A seven point scale eg 0-6, or 1-7 would contain a mid point denoting average performance, ie 3 or 4, with three points for below and three points above for average performance.

In the present study a scale of 1 to 6 seemed appropriate, as it paralleled the scale adopted in question 1B of the English Ability Rating form. A tentative set of definitions or indicators is proposed below:

Level 6. Virtually native-speaker ability.
5. Very competent with some weaknesses.
4. Competent but with weaknesses; nevertheless adequate.
3. Not competent; serious weaknesses; barely adequate.
2. Very weak. Inadequate for study.
1. No ability, or hardly any, to communicate in English.

It will be noted that the word adequate was used in some of the definitions. This was deliberate. It meant that the assessor determined what he considered adequate, barely adequate and inadequate for a 'representative' student, regardless of type and field of study, and that the same criteria were applied to all who were assessed.

Since both the analytic scale and the general scale were incomplete and untried at the time of the enquiry it was decided to use both scales during the trialling of the measure and to make a final decision on the procedure to adopt for the main study and to modify criteria in the light of experience gained during the trialling of the interview.

4.5.4 The dictation.

The purpose of the dictation test was to test the subject's ability to understand information dictated under conditions resembling those

of the classroom and to rewrite that information as accurately as possible. The text to be dictated should be drawn from the type of text that students might expect to meet at university level. It should also be organised in a way that might resemble material that a lecturer would dictate in class.

In 4.5.2.1 above it was suggested that two texts might be dictated, but after further reflection on the time that the measure was likely to take it was decided that one text with 80 items should be sufficient.

4.5.4.1 The topic of the passage was to be decided in the context of the texts and tasks for the other components of the CPM. Since the interview and the writing tests took as their topic areas the immediate interests and experiences of each individual subject, it was only the texts of the dictation and cloze passages which were 'imposed' on the subjects and which therefore had to be taken together. For students following courses in Britain subject areas to be considered are arts, including literature, social science, education, natural sciences, engineering sciences, medicine, agriculture, British culture and more general academic texts eg about university life or work. Suitable sources of texts were course textbooks, academic journals, serious newspapers, books and magazines written for the general educated public but whose subject matter should also appeal to students, and literature of an interdisciplinary nature.

The text finally chosen was taken from the Edinburgh University Bulletin, 1974. The bulletin is published free by the University

for a readership comprising all students, all academic staff and all non-academic staff. It therefore contained articles of a variety of styles, but rarely more than one of a highly specialised nature. The report from which the dictation passage was drawn appeared to be intended for the general reader, although members of the Education Faculty were likely to be most interested in its contents.

4.5.4.2 The presentation of the dictation is not simply a matter of reading the text with pauses at places that seem appropriate. Since the dictation is one part of a communicative proficiency measure it has to be set in that context. The dictation is part of a deliberate act of conveying information to the subjects. They are required to listen and understand, and then at a given point take the rest of the information down in writing.

The dictation therefore had three phases. The first was the introductory phase in which the speaker gave the background and introduced the topic - in this case an account of a primary schools biology scheme. As well as providing the context for the dictation it enabled the listeners to become used to the voice quality of the speaker before the actual dictation. The second phase was the instruction phase in which the speaker paused and indicated that he would be dictating the next section and gave the subjects instructions as to what to do. The third phase was the dictation phase in which the speaker dictated and the subjects wrote. At the end of this phase the speaker either gave the students a minute to look over what they had written or else finished imparting the information on the topic.

4.5.4.3 Since the dictation was to be delivered to groups of subjects in different places it was necessary to ensure that the presentation was the same for all. This meant the text had to be recorded, and the length of the pauses and the length of the 'chunks' had to be determined beforehand. The length of the pauses were worked out by dictating the passage to a proficient writer of English and observing the time he or she took to write down each 'chunk'. The 'chunks' - or sections of the text which are dictated between the pauses - were determined by dividing the passage first into sense groups. Certain limits were then observed. The minimum length of 'chunk' should be three words (three syllables) and the maximum twelve words (eighteen syllables), with an average of about six words (nine syllables) per chunk. If necessary chunks could be longer in the second part of the text. Length of pauses can varied between three and twenty seconds.

4.5.4.4 The introduction to the text chosen for dictation was a resumé of the first part of the report and was ninety-nine words long. The instructions, sixty-six words long, informed the subjects that they had to write down what followed, that no sentences or parts of sentences would be repeated and that no punctuation would be given. They were instructed to supply the punctuation and write any numbers in figures. The speaker then referred back to a key word in the text, developments, and reiterated that he would dictate these developments. He further indicated that the developments were numbered from one to four. The instruction to write was repeated and the dictation proper began.

The version for trialling was recorded and given to two adult native speakers of English first. Their writing was observed by the investigator. As a result of this pre-trial certain chunks were changed and about a third of the pauses reduced in length.

Paragraph 3 of the dictation follows, first as it was originally recorded, and secondly in the revised version. The numbers in the gaps represent the length of the pauses in seconds.

3. It is clear 5 that teachers would like more visual aids 18 in the form of charts, 8 posters, models, slides etc. 8 A small collection of these could be built up 20 if funds were available. 12

It was found that the first three pauses were too long, and they were reduced by 3, 6 and 2 seconds respectively. It was also found that the chunking of the second sentence was not entirely satisfactory. A longer break after these (fifth word) was more natural than a long break after up. Accordingly for the trial version a break of 8 seconds was introduced after these and a pause of 20 seconds at the end of the sentence. The revised paragraph now reads.

> 3. It is clear 2 that teachers would like more visual aids 12 in the form of charts, 6 posters, models, slides etc. 8 A small collection of these 8 could be built up if funds were available. 20

4.5.4.5 Scoring the dictation was a further problem. Traditional methods have been to award as many marks as words and to deduct a mark for every mistake. Specification of eighty items in para 4.5.4 above implied that a decision had already been taken to use a selective system of scoring, adopted by Fountain (1974), Taylor (1975) and others. For this dictation it was decided to score only

those words with lexical load eg project, discussed, successful, together with some key structural words with significant lexical content such as modals and connectors. In the investigator's view these items were the most important for comprehension of the text. The total number of items was rounded to 80 for ease of calculation of results. All items were marked correct or incorrect, with any mistake - however insignificant - rendering an item incorrect. Inaccuracies in the rest of the text were ignored. The words that were scored were underlined in the key. Thus the key to paragraph 3 was as follows:

> 3. It is clear 2 that teachers would like more visual aids 12 in the form of charts, 6 posters, models, slides etc. 8 A small collection of these 8 could be built up if funds were available. 20

Etc was not scored as it was anticipated that some students might write it in full and others in abbreviated form. This would also raise the question of whether the full stop after etc was part of the spelling and so to be included, or the full stop at the end of the sentence. Three structural items were considered important and scored. These is important in that it refers back to the visual aids of the previous sentence; if is important for the meaning of the proposition, and it is important that the listener appreciates the condition imposed; and were is an optional form which a careless listener may not pick up. The paragraph contained 32 words (43 syllables), dictated in 6 chunks with 18 words to be scored.

It was not clear before the trialling what particular scores would mean. But in general terms it seemed reasonable to expect a student whose English was generally adequate for his studies to write the

dictation with at least 80% - 64/80 - accuracy. It was decided to await the outcome of the trialling before making further decisions or interpretation of the scores.

4.5.5 The essay.

The purpose of the essay was to give the subject an opportunity to express himself freely on a familiar topic in writing and so to provide a sample of writing that would be sufficient for an assessor to form some conclusions as to the adequacy of the subject's proficiency in written English. Because the test was one of communicative writing the subject was not be asked to write a 'treatise' or set piece but to give an account of and discuss some aspect of his studies or recent experiences while in this country.

4.5.5.1 Three factors were taken into consideration in the presentation of the writing task. First was the adequacy of the sample of language to be produced. The finished piece of writing had to be of sufficient length and to embrace more than one style if possible so that a judgement could be made with some confidence. Accordingly it was thought advisable to give the writer some indication of the length of the writing and to set a task or tasks that would demand more than one style of writing. The basic options here were either to require the subject to undertake more than one writing task or to combine different subtasks into one general writing task. The two types of writing that seemed most appropriate in the present context were the presentation of some facts or events and some discursive writing setting forth a few simple opinions or comments on those facts.

4.5.5.2 A second factor was that of time. On the one hand enough time had to be given for the subject to complete the task and on the other hand, time was not plentiful if each subject was to complete the CPM. 30 minutes seemed to be the minimum amount of time to be allowed.

4.5.5.3 A third major factor was the topic to be written about. As in the interview all subjects should perform the same type of task but with the subject area related to the subject's own experience. The writing task therefore resembled the third phase of the interview, but in written form, preferably not duplicating what had already been discussed orally. It was also important that the subjects should not have to spend undue time organising the content or wondering exactly what they should write about. Clear guidelines should help alleviate this problem. Such guidelines should offer suggestions as to the kind of facts to include in the first part of the essay and suggestions as to some possible reasons or opinions which might be included in the second part. In addition the subject should know what the test constructor was really looking for. The normal reaction in examination conditions is to be concerned about the accuracy or validity of the contents and about whether penalties will be incurred if the task is not finished within the specified time limit. In the present case neither of these factors applied and the subject was to be made aware that it was not the accuracy of the content that was required most of all but the fluency of expression and the style. Consequently it was not important to have completed the essay, although the second part should have been well on the way to completion.

4.5.5.4 Three ways of assessing the work produced were possible resembling the approaches considered for the interview. The first option was to adopt an analytical approach by assessing individual features of the written work separately. The second was to give an overall impressionistic mark, while the third was to consider the two.

4.5.5.5 If the analytical approach was to be adopted, the first decision to be taken was the identification of the different aspects of the discourse that were going to be assessed individually. The number of aspects chosen was not as critical in the assessment of writing as it was in the assessment of speaking since the assessor was not taking part in live interaction but was able to re-read the written communication as many times as were necessary for his purpose. However, if too many parameters were examined, there could be dangers of overlap. The very broad parameters related to the nature and organisation of the content, the quality of the expression, the appropriateness of the style and the complexity of the sentences written.

The purpose of a communicative writing task is to encourage the writer to express certain information and ideas so that they can be readily comprehended by the reader. Accuracy of usage or of spelling is secondary, provided that any weaknesses in this area do not seriously impede comprehension or distort the message. The relevance and organisation of the content and the comprehensibility of the writing are the first important features to be assessed. Attention can then be turned to the quality of the language used.

In the first place the range and appropriateness of the lexis and syntax used can be assessed. One score could be allotted in respect of syntax and another in respect of lexis. However, where the emphasis is on successful communication, these two features could be combined in one scale. A further indication of the writer's communicative proficiency is the complexity of the sentences that he uses successfully. A final parameter would account for the accuracy of expression - predominantly of the syntax, but including lexical choice and spelling. Persistent deviant usage has an adverse effect on comprehension.

In the context of the present study the following criteria were proposed for an analytical assessment:

i. Relevance, extent and organisation of content.
ii. Range and appropriateness of syntax and lexis used.
iii. Complexity of sentences produced.
iv. Accuracy of expression (syntax, lexis and spelling).

Each of the four criteria could be assessed by a 10 point scale with 1 representing total absence of the criterion and 10 entirely satisfactory use of the criterion. Each would have equal weighting.

4.5.5.6 The alternative approach was to give an overall impressionistic mark on an appropriate scale. This has been the traditional form of essay marking. In some cases definitions are provided for the points on the scale but these are not always precise or satisfactory - see discussion on BCSA in para 3.3.2 above and Table 3.5. The BCSA uses a five point scale, A to E. However, in the context of this study a six point scale parallelling the scale used in question 1B of the English Ability Rating form and the scale proposed for the CPM interview (see 4.5.3.7 above) seemed

the most appropriate. A suggested list of definitions was:

Level 6. Virtually native speaker writing ability.
5. Very competent. A few weaknesses.
4. Competent but with weaknesses, nevertheless adequate.
3. Not competent, with serious weaknesses; barely adequate.
2. Very weak; inadequate for academic writing.
1. No, or hardly any, ability to communicate in written English.

As in the definitions proposed for the interview the word adequate was used for a number of definitions. The adequacy referred to adequacy in English for a 'representative' student in the judgement of the investigator. No attempt was made to take into account the particular field of study of the individual writer.

4.5.5.7 It was decided to adopt both approaches to the scoring during the trialling and to decide which to adopt for the main enquiry in the light of experience in the trialling.

4.5.6 The reading test (cloze).

The purpose of the reading (cloze) test was to gain some indication of the sample's general English proficiency as revealed through reading. While not apparently as functional nor as communicative a task as the tasks of the other tests, completion of the cloze tests required the subject to interact with the text.

The major decisions to be taken with regard to the contents of the cloze test were:

how many passages should there be?
what should the topics be?
how many deletions should be made in each (length)?
what rate of deletion should be adopted?

It has already been suggested that four passages with a total of around 120 deletions would be desirable (see para 4.5.2.2 above), although it was estimated that more time than was available might be required. For the purposes of the trial it seemed nevertheless desirable to find four passages to form the basis of the test.

4.5.6.1 The general topic areas of the passages have already been discussed in the context of the dictation (see para 4.5.4) together with possible sources. Since a more general passage from the field of education had already been chosen for the dictation, priority subject areas for the reading were science, medicine, maths, technology or engineering, social science, perhaps another general academic topic, and a passage which reported research. Sources should be journals, 'popular' science publications, general books in the area and publications such as a university bulletin.

There is no commonly agreed length for cloze tests. Since the original use of cloze procedure was for determining the readability of texts, the question of length did not really arise. However, length increases the reliability of a test and fifty items represents a long enough test to obtain acceptable reliability and would represent a text of sufficient length to treat a particular theme adequately. However, in the present context observing that kind of length would mean effectively only two passages, whereas if the range of passages was to be maintained, 25 to 30 deletions were all that were possible. It was the writer's opinion that 25 represented the minimum desirable number of deletions.

4.5.6.2 The general recommendations of existing literature on cloze procedure are to delete at the rate of every 5th, 6th, 7th or 8th word. At the time of preparing the CPM Alderson had not yet reported any findings on his enquiry into deletion rates (Alderson, 1978). The two major options were to choose either the same rate for all the texts, or to vary the rate in each one, thus giving some variety in the task. However, as it was anticipated that subjects might not complete all the passages or have freedom to choose the passages they wished to tackle, it was advisable to keep the same rate of deletion in order to eliminate a possible extra variable of deletion rate. One further option remained - and it would certainly introduce an extra variable - to delete unsystematically, or selectively. Selective deletion may well alter the nature of the task, since if only syntactic items are deleted, it can be argued that it is a test of syntax, and if lexical items are deleted, the test would be principally a test of lexis.

4.5.6.3 Four passages were chosen for the reading test. Three were written on topics from the general area of science and one in the general area of social sciences. Two passages came from scientific journals, one from the University Bulletin and one from the introduction to a book. Three passages were prefaced by short paragraphs summarising what had gone before and placing the passages in context. One passage contained 31 deletions, reduced to 30 owing to an undetected typing error, another passage 30 deletions, and the remaining two twenty-five deletions. It was anticipated that if any words proved very difficult or impossible to complete, these might be supplied in the finalised versions.

The rate of deletion chosen was every 6th word. This applied to all except one passage - passage MC - where it was decided that only syntactic words should be deleted. This text therefore became a modified cloze passage. There were 25 deletions in the 129 words that followed the introductory paragraph and a 'lead-in' sentence. The average rate of deletion was therefore approximately one deletion for every five words of text. Since the reading test was essentially a general proficiency test it was felt that to make one of the four passages predominantly a test of grammar use, would not distort the validity of the test as an indicator of general proficiency. It was anticipated that subjects might find this particular text a little easier to complete than the others.

The characteristics of each of the passages are given below. The original texts appear in Section 2 of Appendix I.

Text AC	Subject	General Science - discovery of a comet
	Style	Report
	Source	University of Edinburgh Bulletin, December 1973
	Length	Introductory paragraph, plus 186 words
	No of deletions	30
Text BC	Subject	Social Science - discussion of magic
	Style	Academic discussion
	Source	Introduction to Facing Mount Kenya, Jomo Kenyatta, 1938, London, Martin Secker and Warburg
	Length	Introductory paragraph, one complex lead sentence, plus 150 words
	No of deletions	25
Text CC	Subject	Mathematics and education. General
	Style	Narrative and discursive; general academic
	Source	Not recorded
	Length	Two introductory sentences, plus 188 words
	No of deletions	31 (reduced to 30 by typing error)

Text MC	Subject	Sex-control technology. Research report
	Style	Descriptive
	Source	Journal of the American Association for the Advancement of Science, May 1974
	Length	Introductory paragraph, plus lead sentence, plus 129 words
	No of deletions	25

Verbal instructions were given to the students indicating that they should fill each blank with one word, and that contractions, eg, weren't, counted as one word. They were also asked to make a note of the time each passage was started and finished in order to give some guidance as to the time taken.

4.5.6.4 Exact word scoring is the most convenient form of scoring cloze for the assessor. The acceptable word method is more suited to the use of cloze in a teaching situation or where students are to have their scripts returned. This method involves either repeated subjective judgement on the part of the assessor, for each alternative has to be considered carefully, or it means that passages have to be tried out with groups of native speakers to find out what the possibilities are. Darnell has developed this method further by assigning a logarithmic value to each native speaker's response (clozentropy) but studies have shown that where either the acceptable word or clozentropy methods have been compared with the exact word method, correlations have been very high (Anderson, 1971; Pike, 1973). For the purposes of the present study the exact word method was considered adequate.

At the design stage of such a test it is not clear what individual scores, or bands of scores, might mean. In the present enquiry it

was decided to wait until the trialling was completed before attempting to interpret the scores obtained.

4.5.7 Trialling the Communicative Proficiency Measure was planned for the end of the academic year, since two components - the interview and the essay - depended in large measure on the subjects' reporting their experiences while in Britain. It was hoped that individual subjects trialling the measure would be able to take as many components as possible. There were two reasons why this was likely to be difficult to achieve. Firstly students were naturally very busy with revision at the end of the academic year and would not wish to devote much of their time <u>voluntarily</u> to doing language tests. And secondly the test as planned would take up to two hours, and it was most doubtful if many, or any, students would be willing to submit to that amount of testing!

The purpose of the trialling was to establish the suitability of the instruments and to modify the contents and procedures where necessary in the light of the experience gained. There was unlikely to be another chance of trialling the revised measure before the main administration during the second half of the following academic year. It was therefore not essential that all the students completed the test, although most desirable. It was also necessary for at least the cloze procedure passages to be trialled by native speakers as a check that they were not too difficult.

4.5.7.1 As arrangements were being made for the trialling, it became increasingly evident that the time factor was becoming very

important and that even at the pre-trialling stage some modifications would have to be made. Three of the four components of CPM involved writing, and it was decided that the writing tasks should be the focus of the first trial group. Since there was one test - dictation - which had to be started and carried out by all subjects simultaneously, it was decided to set a fixed time for the testing and to begin with the dictation. This could then be followed by the essay and finally by the cloze tests. If any subjects arrived too late for the dictation, it could be readministered at the very end. The focus of the second trial group would be the interview. Since the cloze tests constituted the final component of the first trialling group and therefore ran the risk of not being completed, they were given priority, among the writing tasks for the second trial, followed by the essay and, if time was left, by the dictation.

In the meantime the four cloze passages chosen were given individually to three female native-speaker postgraduate students in the Department of Linguistics at the University of Edinburgh. Each person was asked to make a record of the time taken to complete each passage and to return them to the investigator within 3 days. It was felt that these students would be aware of strategies for doing cloze, and since none of the scientific passages appeared to require technical knowledge, it was not felt that they would be handicapped. It was felt that if this group of students had serious difficulty with any of the passages - ie achieving scores of well under 60% - the passages in question should be dropped. The trialling of the cloze passages took place just before the first trial with non-native speakers.

4.5.7.2 In addition to the small sample of native speakers for the cloze tests, arrangements were made to obtain two samples of non-native speakers - group 2 in Edinburgh and group 3 in Aberdeen. The group in Edinburgh comprised students studying under the auspices of the British Council at the University of Edinburgh and at the Moray House College of Education. Dates, times and rooms were arranged for the testing - on two separate days in the University (Department of Linguistics) and on one occasion in Moray House (Scottish Centre for Education Overseas). The second group was to be tested in Aberdeen and was to consist of a group of Burmese students who were to be tested at the British Council centre there. A letter was written and sent to about 50 students requesting cooperation, stating that the duration would be approximately 1 hour and 15 minutes on a walk-in basis, explaining the context of the research and offering only a cup of coffee, scores and confidentiality in return!

The samples planned for the trialling, together with the subtests to be trialled are summarised as follows:

Group 1. Native speakers of English. (Cloze passages)
Group 2. Non-native speakers from University of Edinburgh and Moray House. (Dictation, essay, cloze)
Group 3. Non-native speakers from University of Aberdeen. (Interview, cloze, plus essay if time)

Three native speakers constituted group 1 - see para 4.5.7.1 above. Only 6 subjects appeared in Edinburgh University, and none volunteered at Moray House! Group 2 were all postgraduate students as follows:

S1	Female	from Poland	studying Applied Linguistics
S2	Male	from Burma	studying Biology
S3	Male	from Poland	studying Geography
S4	Female	from Thailand	studying Education
S5	Male	from Yemen	studying Medicine
S6	Male	from Iraq	studying Biochemistry

Group 3, postgraduate students in Aberdeen, comprised two female and three male Burmese studying Forestry, Petroleum Geology, Botany, Soil Sciences and Forest Pathology.

The numbers were very disappointing but it was possible to trial all the subtests, and the experience pointed up the importance of extensive preliminary negotiations to obtain samples, and the need for departmental and staff cooperation.

4.5.7.3 For the administration to the Edinburgh sample, group 2, the purpose of the testing was explained to those present and then the dictation was administered. The three phases of the dictation had been recorded on tape and the tape was played to the subjects. The quality and the volume were satisfactory. The whole administration took 15 minutes.

A sheet with the essay topic, instructions and guidelines was distributed and the subjects were instructed to spend only 30 minutes on the task. The instructions were followed well and length of essay varied from two thirds of a foolscap sheet to about 1 side and a quarter. The time seemd to be ample, the quickest finishing in 20 minutes. As each subject finished the essay, he was given three cloze passages together with verbal instructions. It was realised that there would not be enough time to complete all four, so 'bundles' had been arranged as follows: AC, BC, MC; BC, CC, MC, and AC, CC, MC. Thus each student was asked to do the modified/ selective cloze (MC) and two others. Only one subject did not have time to do all three and was able to attempt only two passages. The subjects were able to complete three passages in approximately

30 minutes. Thus the actual administration of the tests was accomplished in the one and a quarter hours originally promised.

4.5.7.4 One week elapsed between the first administration to the Edinburgh group and the administration to the Aberdeen group. There was thus time to assess the results of both group 1 and group 2 performance. The results are discussed below (in para 4.5.8 ff), but one finding affected the subtests administered to group 3. There was evidence from the native-speakers that passage BC was too difficult and that a replacement passage was needed. A new passage was found from a non-academic source in the hope that it would prove fairly easy. Its characteristics are:

Text DC	Subject	General interest (social science)
	Style	Narrative, discursive
	Source	'Ladybird Books'
	Length	Introductory paragraph, plus 158 words
	No of deletions	25

Five students arrived at the same time for the test in Aberdeen. However, they did not wish to stay more than an hour and it was clear that there was no question of administering the essay and dictation. The purpose of the testing was explained briefly and the cloze passages were then distributed. In addition to text DC replacing text BC, a new approach was adopted in the light of certain comments made by subjects after the Edinburgh trialling. Subjects were given all the passages and asked to complete them in any order. Three subjects managed to complete all four, one completed three, while another was only able to complete two. The two students who failed to complete all four were the weakest at the interview and were third and fourth to be interviewed. The interviews took longer than expected - from 15 to 24 minutes.

4.5.8 Results of the trialling.

4.5.8.1 The dictation was administered to group 2 as the first part of the communicative proficiency measure. The group found it an easy task, all except one student gaining scores of 64 (80%) or more. The maximum was 80. Individual scores were:

78, 77, 72, 72, 64 and 48. Mean 68.5

Main difficulties occurred in the longer chunks, particularly with the following phrases:

Para 2 - Arrangements for borrowing university material.
Para 3 - Could be built up if funds were available.
Para 4 - Will remain the first year medical or biological classes.

The dictation results are examined further in the context of the full results for the group in para 4.5.8.6.

4.5.8.2 The results for the essay are given in Table 4.5

Table 4.5 Results for the essay test: Trial Group 2

	Analytic					Impression
	C	S/L	C/S	Acc	T	
Max	10	10	10	10	40	6
S1	6	9	9	9	33	5
S2	8	7	7	6	28	4/5
S3	9	7	5	4	25	4
S4	3	6	6	6	21	3/4
S5	5	6	5	3	19	3
S6	6	3	1	-	10	2

C = Relevance, extent and organisation of content
S/L = Range and appropriateness of syntax and lexis
C/S = Complexity of sentences
Acc = Accuracy of expression (see para 4.5.5.5)
T = Total

Interpretations of the impressionistic scale 1-6 are given in para 4.5.5.6.

The impressionistic marks were awarded first in accordance with the definitions given at para 4.5.5.6 above. In the cases of

students S2 and S4 it was not exactly clear which definition was most appropriate. S2 was clearly competent and very communicative (level 5). However it was clear that he had more than a few weaknesses or inaccuracies of expression and could not therefore be rated much above level 4. The /5 indicates that in some aspects he exhibited level 5 ability. Similarly student S4 could not with confidence be rated as level 4, even though the range of vocabulary and accuracy were of that level. But the lack of organisation and the nature of the content bordered on the unacceptable for a university student and so she was awarded level 3 with /4 indicating signs of level 4 performance - in this case linguistic. Only S6 failed to reach the criterion 'barely adequate' (level 3).

The analytic marking was carried out once the impression marking was completed, and without reference to it. Only in the marking for accuracy was some element of objectivity introduced. It was assumed that at least 90% of what was written would be grammatically accurate. Anything less accurate received no credit at all - see S6. This form of marking did produce differences between handling of the content and expression. S1 conveyed very little in her essay but showed almost native-speaker ability in her expression of this non-message! S4 also had little information to communicate, but what she wrote revealed adequate linguistic proficiency. By way of contrast S3 had plenty of interesting information and ideas to impart but his attempts at complex sentences and his inaccuracies of expression did not equal the standard of the content of his writing. S6 was a more extreme case in which there was a very serious attempt to convey some information which was for the most

part expressed in simple, inaccurate language. Subjects S2 and S5 did not reveal such differences in their treatment of the content and their expression.

The assessor had assumed that the stated categories were different. The difference between assessment of the content and its organisation and the other linguistic features was most marked in the case of S1 and S6 who had each been awarded 6/10 for content but ranked 1st and 6th respectively overall. The differences between assessment of accuracy and complexity of sentences used were minimal. Rank ordering of the six on these two scales was almost identical. It could be argued with some justification that there is very little difference between these two features - complexity of sentence and accuracy of syntactic use. The main distinction seems to lie between the content (and its organisation and relevance to the topic given) and the quality of expression. The communicative effect of the writing depends on both these broad characteristics.

No significant difference in rank ordering was evident when the subscores were grouped differently. Table 4.6 below contains total scores achieved in three different ways. First, T is the total accounting for all 4 subscores. C + A is the sum of the scores for content and accuracy, and C+ $\frac{(C/S + A)}{2}$ is the sum of the scores for content and complexity of sentences and accuracy divided by two.

Table 4.6 Essay scores with totals computed in different ways

	T	C + A	$C + \frac{(C/S + A)}{2}$
Max	40	20	20
S1	33	15	15
S2	28	14	14.5
S3	25	13	13.5
S4	21	9	9
S5	19	8	9
S6	10	6	6.5

Totals C + A and $C + \frac{(C/S + A)}{2}$ ignore the feature 'range of syntax and lexis' and reduce the weighting for the assessment of the language. As a result S1, S2 and S3 are grouped much more closely together.

It will be seen from Table 4.5, para 4.5.8.2, that in this small sample there was a 1.0 correlation (rank order) between the two sets of results. This was admittedly a very small sample and too much should not be read into the results. It nevertheless confirmed findings by Pilliner (1974), Ingram (1971) and others that there tends to be a close relationship between analytic and impressionistic marking. Both methods of marking place heavy emphasis on the consistency of the assessor. This was the only option open to the investigator as resources were not available to employ other markers. For the main study it was felt that the impressionistic marking would be sufficient provided that the assessor took into account the relevance and organisation of the content, the complexity of the syntax employed and the accuracy of expression when deciding the level to award. It was also felt that the definitions for each level should be improved.

There were no major problems noted over the administration of the

essay. Oral advice from the invigilator was followed rather than the guidelines that appeared on the question sheet and it was felt that this procedure would be preferable for the main study.

4.5.8.3 Two groups took part initially in the reading test - group 1 and group 2.

Group 1, the native speakers, varied considerably in the amount of time taken to complete the passages. The most rapid reader hurried through the four passages in 26 minutes, while the most painstaking (the highest scorer) took 58 minutes. The modified cloze passage together with passage AC were found to be the easiest with mean scores of 79% and 77% on each, although subjects tended to spend longest time on passage AC. The shortest time (average 6.5 minutes) was spent on passage CC even though it was one of the longest with 30 items. Passage BC was clearly the hardest with a mean of only 48% and an average of 13 minutes spent on the 25 items.

Table 4.7 Results of reading test - Group 1

Student	Test, no of items and scores (time in brackets)				
	AC 30	BC 25	CC 30	MC 25	Total
NS1	24 (20)	14 (16)	22 (7)	20 (15)	80 (58)
NS2	23 (4)	12 (8)	19 (7)	16 (7)	70 (26)
NS3	22 (15)	10 (15)	20 (5)	23 (10)	75 (45)
Means	23 (13m)	12 (13m)	20.3 (6½ m)	19.7 (10½m)	75 (43m)
Means %	77	48	68	79	68

Although the sample was very small the following observations were made:

- The rate at which native speakers complete cloze passages can vary considerably.
- The greatest variation in scores was achieved on the selective cloze test, but this may have been due to the speed and possible carelessness of the particular subject.

- Passage BC took the longest time to complete (half a minute per item on average) and was the most difficult, well below the 60% limit expected of native speaker performance.

- The other passages were satisfactorily completed with all subjects gaining scores of over 60% per passage.

If passage B were not taken into account the average facility for the cloze reading test would be 75% and the average time taken 30 minutes.

Group 2, the non-native speakers at Edinburgh University, attempted the reading test after having completed the dictation and the essay. The trialling was not altogether satisfactory as they were not expected to do all four passages, and even found their timing running out. Half an hour was necessary for those finishing three passages. One student (S5) attempted three passages but was unable to understand the third passage, while S6 did one passage - with the subject matter nearest his own field - but did not attempt any more. All the subjects attempted passage MC but the other passages varied - either AC and BC, BC and CC or AC and CC. Table 4.8 gives the number of students (N) attempting (ie scoring on) each subtest and the mean scores in percentages. Because of pressure of time it was not possible to check individual time taken per passage.

Table 4.8 Results of reading test - Group 2 (N = 6)

Passage		N	Mean	(%)
AC	30	3	9	30
BC	25	3	7.5	30
CC	30	4	11	37
MC	25	5	10.5	42

Although the sample was very small it was clear that the group found the passages about twice as difficult as group 1 did. They found the selective cloze the easiest (MC) but found BC and AC equally difficult.

Because of the difficulty of text BC and the time taken to complete it, it was decided to eliminate it from the battery and substitute another with 25 deletions on a general social science topic, passage DC. (For fuller details see para 4.5.7.4 above). This was trialled with Group 3 in Aberdeen and later by a group of native speakers, students in Moray House.

Group 3 were asked to do all four passages - AC, CC, DC and MC in whichever order they chose. While they were attempting the passages, they were called out individually for interview. This was not a very satisfactory arrangement, but the only one possible given the limited space and time at our disposal. Three of the students finished all four texts, some completed three, and one only had time for two. The two students failing to complete the passages were adjudged to be slightly weaker in their oral proficiency than the other three, and scored proportionately lower on the passages they did choose. All five students chose to do passages CC and DC whose subject matter, maths teaching and the rotation of crops, superficially appeared to be closest to the disciplines they were studying. The selective cloze passage (MC) on sex-control was attempted by only three students.

The group found the passage with MC clearly the easiest. A typing error reduced the number of items in this administration to 24. The group found the other passages of very similar difficulty with performance marginally better on AC. Table 4.9 gives the number of students (N) attempting each passage and the mean scores in raw scores and percentages. Although a total of only 11 non-native students were tested, they nevertheless represented a full range

of ability. Table 4.10 reports the mean scores of each passage with both groups combined. Passage MC is found to be the easiest, but it was not attempted by most of the weaker students.

Table 4.9 Results of reading test - Group 3 (N = 5)

Passage		N	Mean	(%)
AC	30	4	12.7	42
CC	30	5	11.6	39
DC	25	5	9.2	37
MC	24	3	15.7	63

Table 4.10 Results of reading test - Groups 2 and 3 combined (N = 11)

Passage		N	Mean	(%)	
AC	30	7	11.1	37	
BC	25	3	7.5	30	(eliminated)
CC	30	9	11.3	38	
DC	25	5	9.2	37	
MC	25	8	12.4	50	

One further session of trialling was undertaken with a group of 20 native-speaking students at Moray House College of Education who were asked to do passages DC and MC. Once again cooperation was voluntary, in the context of a class on language testing. Regrettably the majority were unable to stay on beyond the end of the scheduled class to complete the second passage. The range of scores (N = 20) on the new passage DC was from 11 to 20, and the mean 15 (60%). It thus remained the most difficult of the passages, although it was taken from a book whose style was supposed to be 'simple' for young readers. Only 7 subjects completed passage MC. Scores ranged from 15 to 22 out of 25 with a mean of 20.5 (82%). This compared with the mean obtained by group 1 of 19.7 (79%).

The results obtained by the smaller sample confirmed the relative

facility of the modified cloze passage (MC) for native speakers but indicated that the new passage DC was more difficult than anticipated. However, it was just within the limit of 60% facility and therefore retained for the main study.

4.5.8.4 The interview was only tried out with group 3, the five Burmese students in Aberdeen. The four phases as described in paras 4.5.3.1 to 4.5.3.4 were followed, although the final phase, questioning by the interviewee, was not always possible. In some cases the second phase, about social experiences in Britain, provided interesting and prolonged discussion whereas in other cases the description of research topics yielded substantial samples of language. The interviewer was in a good position to ask for clarifications and explanations of information as all the subjects were working in disciplines about which he was profoundly ignorant! The warming up, introductory, phases worked well and all candidates seemed reasonably relaxed and cooperative. The very informal and 'homely' atmosphere of the lounge in the British Council where the testing took place was doubtless an important factor. Nevertheless it became apparent that 15 minutes was a minimum period of time, and that frequently 20 minutes were needed if all the phases of the interview were to be covered.

The two sets of criteria discussed in paras 4.5.3.6 and 4.5.3.7 above were applied. Assessments of the seven analytic criteria were made and often recorded as the interview progressed, while the allocation of an overall score was made usually in a brief pause between interviews. Observations made are set out in Table 4.11.

Table 4.11 Results of interview - Group 3

Criteria	Subjects S7	S8	S9	S10	S11
i. interviewee's comprehension	80%	85%	90%	90%	95%
ii. interlocutor's comprehension	80%	90%	80%	90%	95%
iii. communication	Yes	Yes	Yes	Yes	√
iv. fluency	Hesitation	Rapid	Slow	Yes	√
v. lexis	√	√	√	√	√
vi. syntax - accuracy	90%	√	85%	75%	90%
vii. phonology - accuracy	75%	70%	75%	65%	85%
Overall level	5	5/4	4	4	5+

The results were in many ways very tentative. Firstly this limited experience served to test the format of the interview and the practice of certain interviewing techniques such as setting the subject at ease, asking questions which invited longer answers as opposed to Yes or No, interrupting the flow of discourse both spontaneously as well as deliberately, getting subjects to answer questions, and finishing the interview. Secondly it provided valuable practice in assessing the performance observed. No clear scales had been worked out for the analytic criteria except use of the percentage scale to indicate the percentage of comprehension and the percentage accuracy apparently achieved. It was found that awarding percentages was inappropriate to the assessing of communicative effect, range of lexis used and fluency. The first two features were deemed to be either present or absent, and fluency was best described on a cline from Yes (ie very fluent) to hesitant.

It is not easy to assess the level of somebody else's comprehension with any degree of accuracy. Assessments at both i. and ii. tended to be identical, although in one case (S8) the subject seemed to have a little more difficulty in understanding than the interlocutor did, but in another (S9) the interlocutor was experiencing more difficulty. In the latter case there were a number of inaccuracies in the subject's use of both the syntactic and phonological systems. The main gauges of comprehension were the requesting of repetition and the appropriacy of responses to comments or questions.

The third criterion, communication, did not contribute to the overall assessment. It amounted to a re-statement of the first two and in the case of this sample, it was clear that communication was effective, if not always 100%. Quantification was expressed in the comprehension scores.

Fluency was not quantifiable, and so brief notes were made. Fluency tended to be either too slow and deliberate, normal speed but with a number of hesitations, normal speed without hesitations, and rapid or too rapid for easy comprehension. Consequently no 'scores' were given to fluency, only comments.

Appropriateness of lexis was difficult to quantify. The range may not have been wide, but it was very difficult to assess how many more lexical items a subject could have introduced on a given topic, and the number of times that inappropriate lexical items were used was negligible. Again no scores were given, simply a tick to indicate appropriateness.

The accuracy of both the phonological and syntactic usage of the

speakers was more readily quantifiable, although the degrees of accuracy represented by the percentages could not pretend to be accurate. All the subjects were quite careful about their syntax. They avoided syntax of which they were uncertain and consequently committed relatively few identifiable syntactic errors. In questions of pronunciation it was much more difficult to quantify. For a start it is not clear what the norm should be. Candidate S1, for example, with a score of 75% was presumably deviating from the norm in some 25% of his utterances. But what norm is it? The RP speaker or a speaker of educated Scottish English? Or is it a norm of a notional educated non-native speaker of English? In practice this investigator tended to use the last mentioned without being able to define the characteristics of such a speaker or to guarantee the reliability of such judgements. In each case, however, there were features of the subject's speed which did impede comprehension or make it difficult. Deviant stress was noted in the case of all subjects, except S11, where no explanation was recorded as to why his phonological accuracy failed to reach 100%. Serious deviations in intonation were noted for two subjects (S8 and S10) while phonemic inaccuracies were a feature of the speech of S9 and S10.

No total scores were given, nor would any total score be able to account for the communication, fluency and lexis criteria. If the percentages awarded for the comprehension and accuracy criteria were averaged, total scores would be 81, 86, 82, 80 and 91. There was little range in the scores although S11 emerges as the most proficient speaker with the others clustered within six percentage

points. Such overall scores are, however, largely meaningless. The broken down information is more revealing. The range for the comprehension ratings was only 80 to 95, while those for accuracy were 75 to 90 (syntax) and 65 to 85 (phonology). Difficulty with the sound system of English is to be expected from speakers of a language such as Burmese. The fact that their standard was apparently so high was due to the fact that all had had some pre-sessional English, all had been in Britain for nearly one year, and in three cases (S7, S10 and S11) for nearly two years.

The impression markings confirmed firstly that subject S11 was the most proficient - the + having been added to show that the speaker was not far short of the native speaker category - and that the subjects as a whole were of very similar standard (4 and 5). In the case of subject S8, he was awarded a 5 (very competent, with some weaknesses) but with certain characteristics, notably weaknesses of fluency and phonology, that were more appropriate to level 4 performance. It is unlikely that the assessments for both S7 and S8 were particularly valid since they were the first subjects to be interviewed and the interviewer was not yet conversant with the techniques or the standard.

As in the case of the writing it was not always clear which level to award to a subject, even though the definitions were so scanty. There were frequently aspects of a subject's performance which indicated a level above or below the level which was deemed to be dominant.

In spite of the small number in the sample and the lack of spread

of ability, the investigator decided to use the criteria in the analytic procedure as a checklist and to report the overall assessments on the scale of 1-6 with improved definitions in the final study.

4.5.8.5 Table 4.12 presents a summary of the scores obtained on the various components of the proficiency measure. Dictation and reading scores are presented as percentages - the percentage for the reading representing the correct responses as a percentage of the items that each student attempted. Essay and interview assessments are reported according to the level awarded.

Table 4.12 Overall results of trialling of CPM, non-native speakers

Subject	Dict (%)	Reading (%)	Essay	Interview
1	97	53	5	-
2	96	40	4/5	-
3	90	45	4	-
4	90	20	3/4	-
5	80	15	3	-
6	60	9	2	-
7	-	44	-	5
8	-	50	-	5/4
9	-	27	-	4
10	-	24	-	4
11	-	57	-	5+
Mean	85%	35%*	3/4	5/4

(* Group 2 Mean = 30%, Group 3 Mean = 40%)

Table 4.13 Test intercorrelations (rho boosted)

Dictation	-		Reading vs Interview
Reading	.94	-	r = .94
Essay	.99	.97	
	(N = 6)		(N = 5)

In spite of the small sample the very high correlation between performance on different tests was apparent. Thus the performance on the reading test confirmed performance on the dictation, and performance on the essay confirmed the other results. There was a very strong relationship between performance on the reading and on the interview. The cloze tests were very much more discriminating than the dictation test, which was of almost equal length. The dictation, moreover, proved to be an easy test, and since Group 3 were of higher ability than Group 2 (cloze mean of 40% as against 30%), it could be anticipated that they would have obtained scores of 90% or more.

Because of its high facility and high correlation with other tests the dictation appeared to give very little information beyond the fact that most students with essay scores of level 3 (minimum adequacy) and above performed on the dictation with 80% or more success. In Alderson's study it was observed that dictation measured lower order rather than higher order skills, and on longer samples correlations between dictation and cloze ranged from .30 to .82 (Alderson, 1978: 310). Its value to the measure as a whole therefore was in doubt. Two additional practical factors led to the decision to exclude dictation from the final measures. It was clear from the trialling that total administration time had to be reduced. Excluding the dictation would be a first step towards this aim. Secondly, it was the least flexible of the subtests to administer. Unlike the other tests it could not be administered to students individually and required a tape recorder and a room where reasonable quiet could be guaranteed. By eliminating the

dictation it was possible to avoid difficulties that could arise as a result of noise and possible failures of equipment and to limit the task to measures that could be started by students individually at any time and in a variety of circumstances.

The trialling of the cloze passages had highlighted one very difficult passage which was replaced by another which seemed more satisfactory, although still more difficult than desired. But once again time was clearly an important factor. The rate at which people complete cloze passages varies amongst both native and non-native speakers. The non-native speakers preferred doing the passages whose subject matter appealed to them or which seemed relevant. It was clearly impractical to ask every student to complete all the passages as it ignored differences in rate of working. Moreover, the investigator's previous experience in test construction had shown that timing of tests is a factor in discrimination and probably a factor in general proficiency. It was decided therefore to retain all four cloze passages - AC, CC, DC and MC - and to instruct subjects to complete as many of them as they could in a period of 30 minutes. Choice of subject matter, individual rate of working and the time limitations were thus accommodated.

The procedures adopted for the conduct of the interview and the task set for the writing tests worked satisfactorily. It was found that the interview often had to be extended to 20 minutes and that the essay could normally be finished within the 30 minutes stipulated. If students had not finished in that time they would have nevertheless supplied a sufficient sample of their writing for the purposes of the test. While analytic marking had been

undertaken during the trialling its main value had been in giving the assessor some basic criteria, or checklist, to work to, and the six levels with improved definitions were judged satisfactory for the main purpose.

In view of the very small sample, reliability coefficients had not been worked out for the reading test. An estimate of the reliability - split half method - was done for passage DC with the Moray House sample, r = .78. Since the test was short (25 items) and since reliabilities of cloze passages vary, this was considered an adequate indication. Reliability of the interviews and the essays depended on the structuring of the tasks and the consistency of the sole assessor, the investigator, with the aid of the fixed descriptions of the different levels of performance. Adherence to the different phases of the interview for all subjects and the instructions and guidelines given for the essay were judged to be adequate for the reliability of the task. Adherence to the marking checklists and to the descriptions of the performance levels seemed adequate to ensure marker reliability. Where possible 'guest' assessors would be invited to assess subjects' performance according to the same criteria.

The communicative proficiency measure as now revised claimed a certain degree of validity. In terms of construct subjects were given an opportunity to express themselves with a minimum of guidance but at some length both in oral mode and written mode. The structuring of the tasks gave them the opportunity to use a range of lexis and syntax as wide and as complex as they felt able to use. A systematic method of observation and assessment had been

developed. (Greater validity could be claimed if assessment criteria as well developed as the Yardstick scales (1979, unpublished[1]) had been available and used.) In addition a reading test consisting of cloze passages tested reading and general proficiency. In terms of content the topic matter of the reading was taken from literature of the type that students were likely to be called upon to read and the topics of their oral interaction and writing sampled social and academic fields.

4.5.9 The communicative proficiency measure resulting from the trialling was a measure with maximum flexibility of administration with maximum sampling of the language possible in the limited time available. The written components could be completed without full time invigilators, if necessary, and could be started on a walk-in basis as subjects arrived. The time needed to complete the measure was reduced to approximately 1 hr 30 mins. The measure comprised:

Interview	-	fixed structure	20 mins
Essay	-	fixed subject	30 mins
Reading	-	four cloze passages, variety of topics, to be attempted in any order - time limit	30 mins
	-	extra time for administration etc	10 mins
		Total time	90 mins

Copies of the subtests that were trialled together with source texts, where applicable,are at Section 2 of Appendix I, and copies of the communicative proficiency measure with assessment criteria and key are at Section 2 of Appendix III.

---

1. Developed in a small workshop comprising members of ETS (Princeton) Language Testing staff and British Council testing staff held in London, November 1979. Assessment scales for writing and oral interaction were developed.

4.6 Conclusions

Preparatory work for the main study was now completed. Criteria for the selection of the sample had been established and a maximum size set of 1,000 non-native speakers of English per year for two consecutive years. The criterion variable to be assessed was the adequacy of subjects' English for completion of their studies. This adequacy was to be assessed in two ways - by means of tutors' ratings and by an external language test. An English Ability Rating form for use by tutors was developed and trialled, and a Communicative Proficiency Measure comprising an interview, essay and reading tests was developed and trialled for administration to a subsample. Only the English Ability Rating instrument was used in the first investigation (reported and discussed in Chapters 5 and 6) while both instruments were used for the second investigation (reported in Chapter 7).

CHAPTER 5 THE FIRST INVESTIGATION

5.0 The Purpose

The purpose of the first investigation was to try out the English Ability Rating instrument, to establish a framework for analysing the data collected and to formulate some preliminary conclusions. The tutor's rating instrument had already been trialled in respect of its content and administrability and it was now ready for application to the first large scale sample of students. Modifications to the instrument and to the procedures were made for the second investigation where necessary. Because of the considerable logistical problems involved with administering the questionnaire nationally to such a large number of students, it was not possible to complete the trialling of the Communicative Proficiency Measure for inclusion in the first investigation.

5.1 Collection of Background Data

The first task of the academic year was to select the sample and obtain basic background information about each subject. A period of time was spent at the British Council Headquarters in London examining the files of all students administered by the Council who:

(a) had arrived in Britain during the previous 3 months, and

(b) would be spending a minimum of 6 months on their course, and

(c) had had their English proficiency assessed either by EPTB or BCSA before their departure from their home countries.

Any student who did not meet all these conditions was excluded from the sample.

5.1.1 Three types of information were sought for each subject.

The first set of details related to their own home and educational

Figure 5.1 Personal data and home testing results sheet

Name ....
Sex ....
Age ....
Coun ....
Ll ....
Qual ....
....
Univ ....
Dept ....
Degr ....
Leng ....
1 ....
2 ....
3 ....
4 ....
T ....
5 ....
W ....
RE ....
nov 73

background, the second related to their studies in Britain, and the third to the details of their pre-departure English assessment and length of pre-sessional remedial English courses followed, if any. Altogether up to nineteen items of background information were collected for each subject as follows:

i. Name
ii. Sex
iii. Age
iv. Country of residence
v. Mother tongue
vi. Highest educational qualifications obtained
vii. British University or other institution
viii. Department in which studying
ix. Degree or qualification being aimed at
x. Length of course

xi.to Either individual subtest scores on EPTB, or
xiv. individual skills ratings on BCSA
xv. Total score on EPTB or average rating on BCSA
xvi. Part II score, EPTB only
xvii. Rating on oral, if given, EPTB
xviii. Rating for writing, if given, EPTB
xix. Amount of pre-sessional remedial English in weeks, if any.

A form was devised so that this information could be collated in a simple fashion. There was space for data on eight subjects per sheet. A sample of the "Personal data and home testing results" sheet is at Figure 5.1.

5.1.2 The data was collected by the end of November - during the latter half of the first term. Each subject was given an identification number for easy reference and to enable the data to be more easily stored on computer at a later time. The identification number of a male subject consisted of one, two or three digits, eg 4, 78, 321. Female subjects were identified by four digit numbers, the first digit always being a 1. Thus a sample of female identification numbers might be 1009, 1063, 1368. One further factor was taken into consideration - the pre-departure English test taken. If a subject had been given the subjective assessment (BCSA) in his home country 500 was added to the identification number. Thus males who had taken EPTB were given numbers from 1 to 499, and females were given numbers from 1001 to 1499, although the last 3 digits never overlapped with the numbers for male candidates. Similarly the range of numbers for subjects having taken BCSA were 501 or 1501 to 999 or 1999.

5.2 Distribution of the Rating Form

The next task was to classify the subjects according to the university or other institute at which they were studying, so that the rating forms could begin their journey to the individual tutors.

5.2.1 An English Ability Rating Form was prepared for each student. The investigator inserted the student's name, country of origin, university or other institute in Britain and, if known, the subject of study. Each form was accompanied by the key to completing the rating form and a letter for the subject tutor explaining the scope of the study and requesting his or her cooperation (see Section 1, Appendix I for the form, key and letter sent out). The forms and accompanying documents were then sorted according to departments and institutions and geographical areas, where the British Council had responsibility for officially sponsored overseas students. Another letter to each of the British Council regional directors was written requesting them to have their staff distribute the documents to the appropriate tutors. Cooperation from these offices was generous, and all forms were delivered except those for students who had either transferred, gone home or were untraceable for any other reason. There were few of these. A total of 833 forms were distributed in this way.

5.2.2 The sequence of events had been approximately:

November-December	- identification of the sample and collection of background data
November-January	- development and trialling of the rating form
February	- preparation of documents for distribution

March - distribution of forms, and completion by tutors.

Tutors received the documents for the most part by mid March and were requested to complete them by the end of March. The majority of responses were returned during the last week of March and the first few days of April. However, many were not completed until the beginning of the third term and so a few forms were still being returned in early May. 744 forms were finally returned, an 89% response from the tutors, for which the investigator was most appreciative, particularly as no stamped addressed envelope had accompanied the documents.

5.2.3 It was clear that the procedure had been very effective, although inevitably the process from collection of background data to the receipt of all the tutors' ratings had taken somewhat longer than had been hoped - six months. If possible it was hoped to bring forward the whole timetable by up to a month the following year in an effort to get the ratings returned by end March or Easter.

5.3 Characteristics of the Sample

Only details of the subjects for whom tutor's ratings forms were completed were analysed. The sample comprised 833 students from 67 countries with an average age of 30 years. The majority were already holders of first degrees or professional qualifications obtained at the post-secondary level. They were studying at 138 different institutions in Britain, for the most part at Master degree or professional or academic diploma level over the full range of subjects, with 30% engaged in academic or professional studies in the area of the social sciences.

439 had been tested by means of the EPTB before taking up their studies in Britain (mean total score 38.9), and 394 had been assessed by the BCSA (mean rating B+). Just over half the sample had had remedial English tuition for an average of 8 weeks before beginning their studies. These characteristics are discussed in further detail below.

5.3.1 The ages of the subjects ranged from 18 to 49 with one 54 year old. Summary details are given in Table 5.2 below.

Table 5.2 Distribution of age of the sample

Age	Frequency	Frequency (%)	Cumulative Frequency (%)
18-20	11	1.3	1.3
21-25	142	17.2	18.5
26-30	349	42.3	60.8
31-35	221	26.8	87.6
36-40	73	8.9	96.5
41-45	23	2.8	99.3
46-50	5	0.6	99.9
51-55	1	0.1	100.0
	825	100.0	
missing cases	8		
mean	29.9 years		
mode	27		

The sample covered a wide range of ages, with 95% within the range 21 to 40. Just over two-fifths of the sample were concentrated in the 26 to 30 year age range. But in general the sample was considered to be one of mature students. The implications for language learning or language improvement were that the older students - over 30 - with a relatively poor background in English might well have found improvement in English difficult if they had not maintained close contact with the language or if they had not practised second language learning strategies for a number of years.

5.3.2 Subjects came from 67 countries in Europe, Asia, Africa and South and Central America, with the latter grouping being represented by 30% of the sample. The next largest group (24.7%) came from South, South East and East Asia with Thailand, Indonesia, Burma and Nepal providing over 70% of that group. 23.4% of the sample came from the Middle Eastern and North African countries, including the Sudan. While Europe was represented by 19 countries half of the cases came from Turkey (39), Poland (19) and Germany (15).

The sample could not be said to be representative of overseas students in Britain, since it excluded native-speaking English students and students for whom English is a second language. Thus Commonwealth countries such as India, Malaysia, Hong Kong and most anglophone African countries were not represented in the sample because it was not the practice at the time of this survey for the British Council to assess the English of students from those countries coming to Britain for further study. Only five Commonwealth countries were represented in the sample - Malawi, Tanzania, Bangladesh, Cyprus and Malta - by 14 students, less than 2% of the total sample. In the latter part of the 1970s, however, it had become British Council practice to extend the assessment of English in these countries and in other countries such as Malaysia and Botswana.

Table 5.3 gives the distribution of the sample by geographic area. Table 5.4 gives the distribution for 13 countries who provided almost 60% of the subjects of the sample.

Table 5.3 Distribution of sample by geographical area

No of countries	Areas	Frequency	Frequency (%)	Frequency (%)
6	Africa (South of the Sahara)	35	4.2	4.2
15	Middle East & North Africa (including Sudan)	195	23.4	27.6
11	South, South East and East Asia	206	24.7	52.3
16	Latin and Central America	250	30.0	82.3
13	North and South West Europe	65	7.8	90.2
6	East and South East Europe	82	9.8	100.0
67		833	100.0	

Table 5.4 Distribution of 13 largest country groups in sample

Country	Frequency	As % of total sample	Cumulative Frequency (%)
Mexico	76	9.1	9.1
Thailand	56	6.7	15.8
Sudan	49	5.9	21.7
Indonesia	41	4.9	26.6
Brazil	40	4.8	31.4
Turkey	39	4.7	36.1
Chile	37	4.4	40.5
Iran	29	3.5	44.0
Egypt	27	3.2	47.2
Burma	27	3.2	50.4
Algeria	25	3.0	53.4
Nepal	24	2.9	56.3
Peru	24	2.9	59.2
	494	59.2	

212 subjects - 25.4% of the total sample came from 16 Spanish-speaking countries, while 159 - 19.1% of the total sample - came from 13 Arabic-speaking countries. Fuller details are given in Section 1, Appendix II.

5.3.3 Distribution according to sex was male 687, female 146.

5.3.4 Almost two thirds of the sample had already obtained first degrees or equivalent professional qualifications, eg qualified teacher, qualified nurse. Nearly a quarter of the sample had obtained a higher degree eg a Masters degree. Included in this category were qualified medical practitioners. A small number had either completed research degrees (PhD) or had undertaken research at that or at higher levels. The educational background of the sample is summarised in Table 5.5.

Table 5.5 Educational background

	Education completed	Frequency	Frequency (%)
1.	Primary only	0	0
2.	Secondary or equivalent	69	8.3
3.	First degree or professional qualification	529	63.7
4.	Second degree (including MD)	186	22.4
5.	Research degree or research	33	4.0
6.	Diploma	13	1.6
	Total	830	100.0
	Missing data	3	

5.3.5 Over a quarter of the sample were studying in colleges or polytechnics in the greater London area. Other cities where 4% or more of the sample were studying were Edinburgh, Leeds, Reading, Oxford, Manchester, Glasgow and Cambridge.

5.3.5.1 The subjects of study were broken down into ten main areas following the categories used by the British Council in their publications reporting statistics for overseas students in Britain (British Council, 1975). Table 5.6 gives the distribution of the sample according to general fields of study. Students studying

nursing, psychology, psychiatry were all included under category 3, Medicine. Category 5, Education, included students studying courses in the Teaching of English as a Foreign Language, but students studying Linguistics or Applied Linguistics, often including TEFL components, were grouped in category 7 - English and (Applied) Linguistics. This latter category also included a few students who were in Britain exclusively for English improvement courses. It will be seen that the largest group of students was undertaking studies in the general area of the social sciences. Students following academic courses leading to degrees or carrying out research in the social sciences were assigned to category 9 - social sciences (academic), while students assigned to category 8 were in the main following courses leading to Diplomas, or no formal qualification, aimed at professional improvement, eg development administration, urban and regional planning.

Table 5.6 Distribution of sample according to subject area of study

	Subject area	Frequency	Frequency (%)
1.	Agricultural, veterinary sciences	70	8.4
2.	Arts	27	3.2
3.	Medicine, nursing, psychology	91	10.9
4.	Physical and biological sciences	121	14.5
5.	Education, including TEFL	98	11.8
6.	Engineering and technology	140	16.8
7.	English studies, (applied) linguistics	52	6.3
8.	Business, social sciences (professional)	124	14.9
9.	Social sciences (academic)	103	12.4
10.	Miscellaneous	6	0.7
	Total	832	100.0
	Missing data	1	

5.3.5.2 The largest single group of students was studying at the

level of postgraduate or professional Diploma or Certificate without degree status. A quarter of the sample was studying for Masters degrees by tuition, often with a short 3 month dissertation, while just over a fifth (22.8%) was working for research degrees at Masters, doctoral or post-doctoral levels. Only 29 (6.6%) were studying for first degrees. A few students were on practical attachments, eg to the Central Electricity Generating Board or to firms, while others were attached to universities for a few months to update themselves in their field or to pursue non-research study of their own. The distribution by levels of study is summarised in Table 5.7 The same categories have been used in this table as for Table 5.5 (Educational background).

Table 5.7 Distribution of sample according to level of study

	Level of study	Frequency	Frequency (%)
1.	-	-	-
2.	Pre-university	4	0.5
3.	First degree	39	4.7
4.	Masters degree by tuition	215	25.8
5.	Research, research degree	190	22.8
6.	PG or professional diploma/ certificate	320	38.5
7.	Academic attachment	42	5.0
8.	Practical attachment	22	2.6
	Total	832	100.0
	Missing data	1	

5.3.6 439 subjects in the sample were tested by EPTB before leaving their countries, while 394 underwent the subjective assessment (BCSA).

5.3.6.1 The sub-sample tested by EPTB performed as a group slightly less satisfactorily than the original survey sample, with a mean

of 38.9 compared with the standard mean of 40.0 for Part 1. The standard deviation was 4.48 compared with 6.0 for the original sample. However, the combination of sampling differences, use of different test forms, and the fact that the very poor achievers should, theoretically, have been excluded, may have contributed to the small drop in mean score and discrimination. It could also be that students in the mid 1970s were reaching Britain with less proficiency in English than their predecessors of the early 1960s.

190 students (43.3% of the sub-sample) also took Part 2, the speed reading test. The mean was 76 (compared with 70 for Form A) and scores ranged from 8 to 178.

The majority of the sample took Form B. Some may have taken Form A, whereas Form C was not yet available. Table 5.8 summarises the means and standard deviations of the performance of the present sample. Standard deviations were consistently lower than 2.0 in all sub-tests. The mean for those taking Part 2 was higher than for the original sample. In two cases, subjects did not complete the Battery and so no total scores were available. In other cases sub-test scores were not all reported and in other cases only total Part 1 scores were reported.

Table 5.8 EPTB: Obtained means and SD's for the sample

	Test	Sample Mean	Sample SD		Standard Mean	Standard SD
1.	Phonemic discrimination	9.3	1.4	(N = 394)	10.0	2.0
2.	Intonation	10.1	1.8	(N = 394)	10.0	2.0
3.	Reading comprehension	9.7	1.5	(N = 395)	10.0	2.0
4.	Grammar	9.6	1.6	(N = 395)	10.0	2.0
Part 1.	Total	38.9	4.5	(N = 437)	40.0	6.0
Part 2.	Reading speed	76.0	28.2	(N = 190)	70.0	34.0

Total Part 1 scores can be summarised according to the banding adopted for the interpretation of scores - see Table 3.1 and para 3.1.5 above. Students with scores below 34.0 are normally advised that their English is inadequate and that they should follow a minimum of 6 months full time English tuition. They are not normally expected to come to Britain for study. Nevertheless, no fewer than 49, just over 11% of this sample, came to Britain, and none was reported as having had more than 3 months remedial English before joining an academic course. For the purposes of the investigation, this was satisfactory, since ideally the performance of students whose English had been predicted as inadequate should also be investigated. The sample now contained such a group. Table 5.9 gives the distribution according to recommended interpretation of Part 1 total scores.

Table 5.9 EPTB: Part 1 total scores

Part 1 total	Interpretation	N	%
Below 34.0	English inadequate	49	11.2
34.0-39.9	Preliminary tuition needed	223	51.0
40.0-45.9	English should be adequate	136	31.2
46.0 and above	Unquestionably adequate[1]	29	6.6
		437	100.0

No sub-tests of the productive skills are built in to EPTB but assessors are advised to supplement the information given by the Battery with a writing task and an interview, where possible.

---

1. This interpretation is not given in the interpretative information on ETPB but is the investigator's own for the purposes of the present study.

Reported grades for writing tests were given for 176 subjects (40.1%), and grades for oral tests were reported for 174 (39.6%). For the writing, just over half the sub-sample (91) were awarded grades of B or B+ while only 25 were awarded A. In the oral, however, 78 were awarded B or B+ while a further 70 were awarded A. Thus assessors were awarding up to half a grade higher for oral performance than for written production. Grades are summarised in Table 5.10.

Table 5.10 EPTB: Grades awarded for writing and oral tests

	Writing		Oral	
Grade	N	%	N	%
D	1	0.6	1	0.6
C	53	30.1	18	10.3
C+	6	3.4	7	4.0
B	79	44.9	71	40.8
B+	12	6.8	7	4.0
A	25	14.2	70	40.2
	176	100.0	174	100.0
Mean	B(-)		B+	

5.3.6.2 The sub-sample for whom subjective grades were reported amounted to 394. Individual grades for each of the four language skills were reported, but no overall grade. For the purposes of this study a 'notional' overall grade was computed by assigning numerical values to each letter, summing the values for each skill and taking the mean. In not all cases have all four skills been graded. In such cases no overall grade has been computed. In other cases no individual skills have been graded, but simply an overall A, B C or D given. Such cases have been accepted for the study and accounted for in the analysis of the overall grade. Over 90% of

the grades awarded for the single skills were unmodified, ie no plus or minus signs were added. In over 60% of the cases assesssors had given different grades for at least one of the four skills. Thus when the 'notional' overall grades were computed over half were at intervening points between whole and half grades. They have been rounded to the nearest half grade in the tables below.

The mode for the listening, speaking and reading skills was A, whereas for the writing skill it was B. Assessors differentiated further by awarding more Bs and fewer As for speaking than for listening and reading. Grades for individual skills and further computed totals are summarised in Table 5.11.

Table 5.11 BCSA: Summary of grades awarded

Grade	Skill									
	Listening		Speaking		Reading		Writing		Total	
	N	(%)	N	(%)	N	(%)	N	(%)	N	(%)
D	2	(0.5)	2	(0.5)	1	(0.3)	3	(0.8)	1	(0.3)
C	10	(2.7)	22	(5.9)	30	(8.2)	52	(14.2)	11	(3.0)
C+	5	(1.4)	4	(1.1)	4	(1.1)	4	(1.1)	41	(10.9)
B	76	(20.5)	145	(39.1)	89	(24.3)	151	(41.3)	66	(17.6)
B+	14	(3.8)	23	(6.2)	14	(3.8)	23	(6.3)	134	(35.7)
A	263	(71.1)	175	(47.2)	228	(62.3)	133	(36.3)	122	(32.5)
	370	(100.0)	371	(100.0)	366	(100.0)	366	(100.0)	375	(100.0)
Mean	A-		B+		B+		B		B+	

5.3.7 None of the BCSA sub-sample had attempted any section of EPTB, but two fifths of the EPTB sub-sample had been assessed on writing tasks and on interview using the same grading system as for BCSA. It can be seen from Table 5.12 below that the students in the BCSA sub-sample did generally slightly better in the oral and considerably better in the writing. Although mean grades for the oral were B+

for both groups, when the numerical equivalents were examined the EPTB group was at the low end of the B+ band (3.25 to 3.74) with a mean of 3.28 and the BCSA group nearer the centre with 3.42. For the writing both groups had mean grades at B, but this time at the extremities of the band (2.75 to 3.24). The BCSA group had a mean of 3.23, while the EPTB group had a mean of 2.84 - a difference of nearly half a grade.

Table 5.12 EPTB and BCSA: comparison of grades for oral and writing

	Writing				Oral			
	EPTB		BCSA		EPTB		BCSA	
Grade	N	(%)	N	(%)	N	(%)	N	(%)
D	1	(0.6)	3	(0.8)	1	(0.6)	2	(0.5)
C	53	(30.1)	52	(14.2)	18	(10.3)	22	(5.9)
C+	6	(3.4)	4	(1.1)	7	(4.0)	4	(1.1)
B	79	(44.9)	151	(41.3)	71	(40.8)	145	(39.1)
B+	12	(6.8)	23	(6.3)	7	(4.0)	23	(6.2)
A	25	(14.2)	133	(36.3)	70	(40.2)	175	(47.2)
	176	(100.0)	366	(100.0)	174	(100.0)	371	(100.0)

5.3.8 More than half the sample received some pre-sessional remedial tuition in English on arrival in Britain. The average period was for about 8 weeks, but ranged from 2 to 32. Students obtaining EPTB scores of less than 40 and students gaining a B in one or more skills on the BCSA would normally be advised by the British Council to follow a remedial English course. As a rough guide the EPTB and BCSA results reported above led us to expect that 272 candidates who had taken EPTB (223 + 49 in Table 5.9) and 210 who had taken BSCA (ie all those gaining B or less in writing in Table 5.12) - a total of 482 - would have been given some remedial tuition. 479 (57,5% of the sample) did follow such

courses, with a higher proportion from the EPTB sub-sample than from the BCSA sub-sample. Evidence supplied by the comparison of grades obtained on orals and writing, in the previous paragraph, suggested that the general proficiency of the sub-sample who had taken BCSA was slightly greater than that of the EPTB sub-sample. The remedial tuition figures also support that conclusion. A breakdown is given in Table 5.13 below.

Table 5.13 Summary of numbers taking remedial English

Group	N	% sub-sample	% total sample
EPTB + remedial English	264	60.1	31.7
BCSA + remedial English	215	54.6	25.8
Total + remedial English	479	-	57.5
No remedial English	354	-	42.5
	833		100.0

Table 5.14 Length of remedial English tuition

	EPTB		BCSA		Total	
No of weeks	N	%	N	%	N	%
2-4	89	33.7	80	37.2	169	35.3
5-8	84	31.9	70	32.6	154	32.1
9-12	91	34.4	57	26.5	148	30.9
24-32	-	-	8	3.7	8	1.7
	264	100.0	215	100.0	479	100.0

5.3.8.1 The length of periods of remedial English tuition is given in Table 5.14. The most frequent periods were 3, 4, 8, 10, or 12 weeks. This reflected a combination of British Council policy, the lengths of courses available, and the availability of the students. The figures give evidence of the fact that only in exceptional cases were students offered more than 3 months English tuition. In the sample being studied the 8 exceptional cases were

given periods of 24 to 32 weeks full time tuition and came to Britain after having been assessed by BCSA. Subjects who had obtained EPTB Part 1 scores of less than 34 and who came to Britain were still subject to the limit of 12 weeks. Subjects from certain countries, eg Indonesia, Thailand, were usually given at least 2 weeks tuition regardless of their English assessments in order to help them settle down more easily in Britain. It is nevertheless noticeable that leaving aside the 8 exceptional cases, students who had done the EPTB were more evenly distributed with respect to length of tuition with the largest single group following courses of 9 to 12 weeks (34.4%), while the largest single group of the BCSA sub-sample had only 2 to 4 weeks tuition (37.2%).

5.3.9 Once the data obtained during the first investigation was analysed it became clear that three sets of analyses would have to be undertaken - one for the whole sample, one for the EPTB sub-sample and one for the BCSA sub-sample. It was therefore important to establish whether there were any important differences in background characteristics of the different samples.

5.3.9.1 The most striking difference was found in the countries of origin and the regions in which they were situated. The two regions which sent the smallest numbers of students with English assessments were Africa (South of the Sahara) and North and West Europe. Of the 96 subjects coming from 19 countries in those regions only 4 had taken EPTB. Thus almost a quarter of the BCSA sample came from those two regions alone. In three major regions of the world, between two thirds and three quarters of the regional

samples had been tested by EPTB. These regions were S and E Asia (77.7% of the regional sample), E and SE Europe (70.7% of the regional sample) and Mid East and N Africa (62.1% of the regional sample). Only four countries used EPTB exclusively - Nepal (N=24), Poland (N=19), Japan (N=16), and Tanzania (N=3). Only 21 of the 67 countries used EPTB, but it was noticeable that these countries had for the most part tested large numbers.

Over half the EPTB sample (239 or 54%) came from five countries - Mexico (N=69), Thailand (N=55), Indonesia (N=39), Sudan (N=41) and Turkey (N=35). Only three countries used BCSA on more than 20 students - Brazil (N=39), Chile (N=37) and Ethiopia (N=22) out of 63 countries using this form of assessment. Most countries had used one method of assessment, handling only occasional students with the other form. Peru, exceptionally used both methods, testing 14 with EPTB and 10 with BCSA. Fuller details of the distribution are given in Section 2, Appendix II.

5.3.9.2 In terms of age and level of education completed the distributions were almost identical. The mean ages were 29.96 (EPTB) and 29.79 (BCSA). Almost two thirds of each sub-sample had completed first degrees or their equivalents, and a further 22% of each had completed second degrees.

5.3.9.3 There were some differences in subjects being studied and levels at which studies were being undertaken. There were proportionately 3.5% more students who had taken BCSA studying Arts, English and Linguistics and 6.1% more taking Education and TEFL

courses, while students who had taken EPTB were proportionately more numerous in Engineering and Technology (by 3.5%), in Physical and Biological Sciences (by 4.4%) and in the Social Sciences - academic (by 6.6%). 30.5% of those who had taken EPTB were studying for Masters degrees compared with only 20.6% of those who had taken BCSA. Proportionately more of the BCSA sub-sample were studying for research degrees and postgraduate or professional diplomas.

The form of assessment adopted, however, was generally a matter of British Council policy in each country concerned and had little to do with other variables. The main point about the background variables of the two sub-samples was the near identity of the age and level of education distributions, and the fact that subjects from countries sending the largest contingents of students through the British Council had been assessed by EPTB.

5.4 Analysis of the English Ability Ratings

744 rating forms were returned to the investigator either fully completed or partially completed. This represented a most satisfactory 89% response. As long as at least one question, or one part question, was answered the form was accepted, and its data contributed to the analyses. About 15% of the rating forms for subjects whose English had been assessed by EPTB were not returned, while for subjects who had been assessed by BCSA only 7% of the forms were not returned.

5.4.1 The first question contained two sets of responses to the general question:

> "Is the general ability in English of the student (i.e. present ability, not potential) adequate for undertaking specialised studies or research in his/her field of study?"

The first set of responses are summarised in Table 5.15 and the second set in Table 5.16.

Table 5.15 Responses to rating form question 1A

	Responses	N	%
1.	Not adequate	30	4.1
2.	just adequate	312	42.3
3.	completely adequate	395	53.6
		737	100.0
	Missing	96	(11.5% of total sample)

More than half the sample were considered by their tutors to have proficiency in English which was completely adequate for their studies, while only 4.1% were considered to have ability in English which was not adequate. These judgements were confirmed by the responses at Section B.

Table 5.16 Responses to rating form question 1B

	Responses	N	%
1.	Well below satisfactory	3	0.4
2.	Considerable deficiencies a handicap	30	4.1
3.	Many weaknesses but just adequate	137	18.6
4.	Many mistakes, minor handicap	172	23.3
5.	Minor faults but adequate	345	46.7
6.	N-S ability	51	6.9
		738	100.0
	Missing	96	(11.4% of total)

When responses 1 and 2, 3 and 4, 5 and 6 were grouped together, the totals were 33, 309 and 396. These compared very closely with section A responses of 30, 312 and 395. Further analysis of responses showed that in only one case was a student deemed 'not adequate' at 1A and rated in 1B at 3 (just adequate). 14 tutors rated students at A as just adequate but then rated them in

categories 5 or 6 at B. A further 13 who had rated students as 'completely adequate' at A, subsequently rated them in category 4 at B - many mistakes, minor handicap.

A product moment (Pearson) correlation was computed between the two sections at .85. The conclusion drawn was that section B responses corroborated the answers at A but did not add significantly new information on the candidates. It could be suggested, however, on the basis of a few responses that for some teachers 'completely adequate ability in English' embraced many mistakes or inaccuracies that presented the student with a minor handicap.

Whereas 11.2% of the EPTB sample was deemed to have totally inadequate English before leaving their countries and 3.3% of the BCSA sample averaged overall C's and D's, the inadequacy rate as viewed by tutors was not so great.

5.4.1.1 There were variations in the performances of the EPTB and BCSA sub-samples. Proportionately more of the BCSA sample were judged to have inadequate English, but the group with completely adequate English was proportionately larger than for the total sample, at 60.2%. The biggest single group in the EPTB sub-sample (49.5%) were adjudged to be 'just adequate'. Distributions are compared in Table 5.17.

Table 5.17 Distributions of responses to question 1A compared

	Response	EPTB %	BCSA %	Total %
1.	Not adequate	3.2	4.8	4.1
2.	just adequate	49.5	35.0	42.3
3.	completely adequate	47.3	60.2	53.6
		N = 372	N = 365	N = 737
		Missing 67	Missing 29	Missing 96

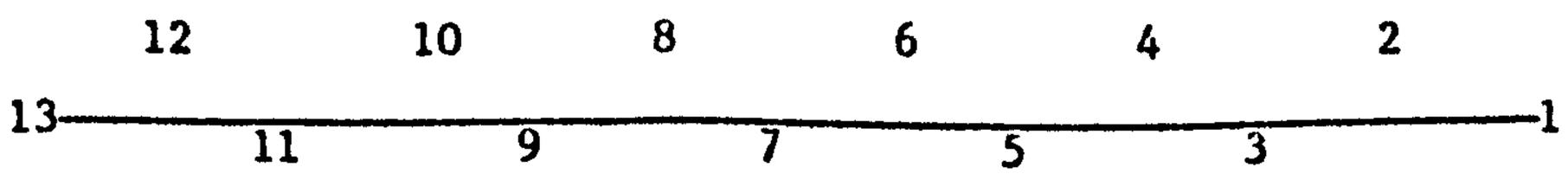
5.4.2 In question 2 tutors were asked to note the student's language skills individually. The entire question is reproduced below as it was on the rating form. Tutors were asked to put X at an appropriate point along each scale.

2. Individual language skills
(Put X at an appropriate point along the scale for each skill)

	completely adequate	adequate	inadequate
Ability to understand spoken English			
Ability to speak English			
Ability to understand written English			
Ability to write English			

The length of each line on the form was 120 millimetres. There were thus 12 sections to each scale with the extreme right represented by 1 and the extreme left by 13. Each point on the scale was 1 cm from the next. This was not revealed to the tutors but was used by the investigator. It made the 'scoring' of the responses in question 2 objective. The scale (see diagram 5.18 below) was applied to each line and the score was simply read off and written on the right of the paper. When the X was placed in the upper half of the centimetre section, a half point was awarded - eg X between 6.5 and 7 was scored as 6.5.

Diagram 5.18 Scale for scoring question 2



5.4.2.1 The complete range of the scales was used by the tutors, even points 1 and 1.5 being used in two cases. However, distributions on all scales were bimodal. The mode for all the skills was 12 with 11.5 and 12.5 also frequently used. The second mode was in all cases 6.5 with other ratings clustered at 6.0 and 7.0. This would seem to indicate that tutors were for the most part deciding either that a subject's ability in a particular skill was just adequate and putting the X somewhere between points 6 and 7 - ie just to the right of the word adequate on the form - or deciding that a subject's ability was entirely satisfactory and putting the X well towards the left end of the line - ie between points 11 and 13.

Over 80% of the tutors distinguished between abilities in different skills. In some cases ability in only one skill, eg speaking, was considered to be different from ability in the others, but in most cases distinction was made between the receptive skills and the productive skills. Further distinctions were made between ability in speaking English and in writing English, as shown in Table 5.19. With such a large sample the differences in means are significant at the 5 percent level.

Table 5.19 Summary of ratings of students language skills (Question 2)

Ratings Scale points	Listen N	Speak N	Read N	Write N
1-3	26	42	12	56
3.5-5	24	58	13	53
5.5-7	250	312	217	316
7.5-9	39	33	39	36
9.5-11	56	49	58	59
11.5-13	344	243	393	196
Total N	739	737	732	716
Mean rating	9.2	8.2	9.7	7.9
SD	3.0	3.1	2.8	3.1

The assumption that appears to have been made by the tutors is that students' reading ability is the most satisfactory, closely followed by their listening ability. If all scores of 5.5 and above are taken to mean at least just adequate, then only 25 (3.4%) of the total sample were estimated to have less than adequate reading ability. The figure doubled again for listening and doubled yet again for speaking and writing.

5.4.2.2 Two interpretations of this pattern of reaction seem possible. Firstly, the order in which tutors have ranked abilities in the different skills represents the order of least to most evidence on which to base an assessment. A tutor has very little evidence of students' reading ability. He can ascertain from students how much of the recommended reading they have completed, and he can infer from their spoken contributions and written work whether certain reading has taken place and been understood. Evidence of listening ability can be gathered either in direct communication or when questions are put to students during class. The tutor can note the rapidity or the appropriateness of the response, or the lack of these. But the opportunities may not be very frequent. Similarly a tutor has samples of a student's speech to assess when he is speaking with a student either in tutorial, in class or informally. Once again the number of occasions on which a tutor has the opportuniuty to speak to a student - even in the first two terms - may be limited. In contrast, writing is the one activity which can be assessed with greater certainty, provided that students have been required to submit written work, because the student's language is recorded on paper and a tutor can re-read it a number

of times if so desired. The fact that fewer students were rated for writing may indicate that in a number of cases insufficient, or no, written work had been demanded of the students.

The second interpretation is that non-native English speaking students, who are reasonably well prepared academically, do indeed experience less difficulty with reading and listening than with speaking and reading. The difference in means for reading and writing is 1.8 points or 15% of the rating scale. Even the difference of 1 full point (nearly 10% of the scale) in the means for listening and speaking is substantial.

It should not be assumed that tutors have given their assessments without any reference to their overseas students. The lack of a self-assessment by students is a disadvantage at this point in the study, as it would have been useful to see the extent to which their assessment of their abilities in English matched their tutors'. Previous evidence quoted in this study (Morris, op cit) and the pre-departure English assessments (see Tables 5.10 and 5.11 above) point to students' basic difficulties with writing.

Both interpretations seem to contain much truth, and assessments were possibly made in the light of both these factors. If the tutors' assessments were therefore valid, the results confirm the notion that students find most difficulty with writing, already observed in the pre-departure assessments, but point to less ability in speaking English than the earlier assessments had led the investigator to assume. The tutors' ratings place reading as the skill in which the students are most able, whereas the BCSA placed listening as the best skill.

5.4.2.3 The four ratings given by the tutors were totalled by the investigator to represent a notional proficiency rating total. The profile obtained from the ratings in respect of each skill was a more meaningful expression of individual students' ability, but a total, assuming equal weighting for each skill and with parallel scales, represented an indicator of ability in English which could be used for comparisons with other total values, such as total Part 1 EPTB scores.

The distribution of the individual skill ratings was bimodal, but the distribution of the total ratings was much less so. The number of cases rose rapidly to the mode at between 5.5 and 7. Thereafter the distribution flattened out to a level of around 80% of that of the mode, rising finally to between 11.5 and 13. Table 5.20 summarises the distribution of 'notional' total ratings for the whole sample as well as for the EPTB and BCSA sub-samples. The distribution for the EPTB was unimodal with the mode at between 5.5 and 7 whereas the BCSA sub-sample distribution followed the total sample distribution a little more closely with the mode at between 11.5 and 13. The distributions are compared in diagram 5.21 while distributions for the listening and for the notional total are compared in Diagram 5.22.

Table 5.20 Summary of ratings for 'notional' total proficiency

Ratings Scale points	All N	EPTB N	BCSA N
1-3	12	5	7
3.5-5	56	30	26
5.5-7	182	103	79
7.5-9	147	82	65
9.5-11	142	69	73
11.5-13	173	66	107
Total N	712	355	357
Mean rating	8.8	8.4	9.1
SD	2.6	2.5	2.6

Diagram 5.21 Distributions of ratings for 'notional' total proficiency compared

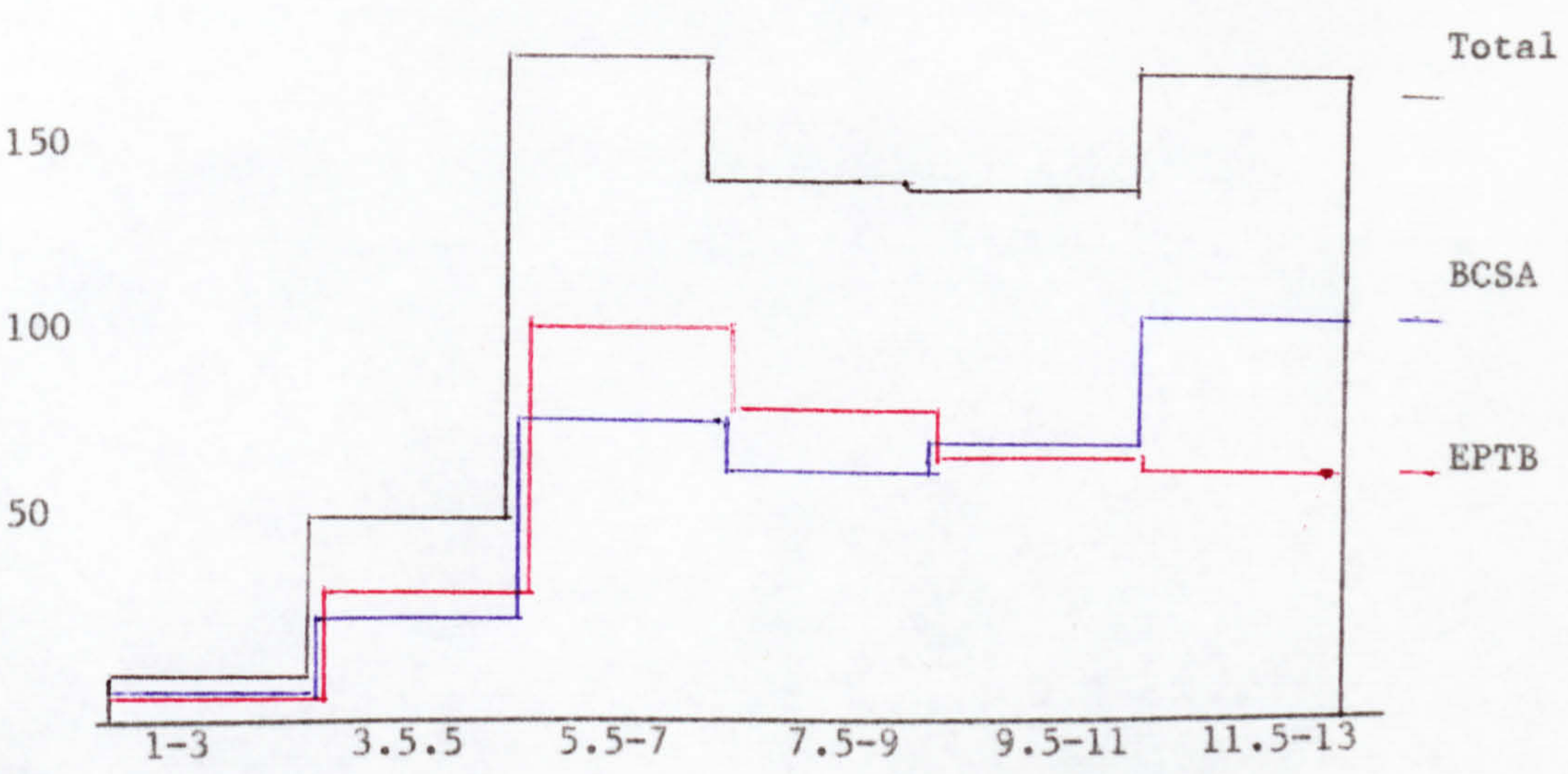
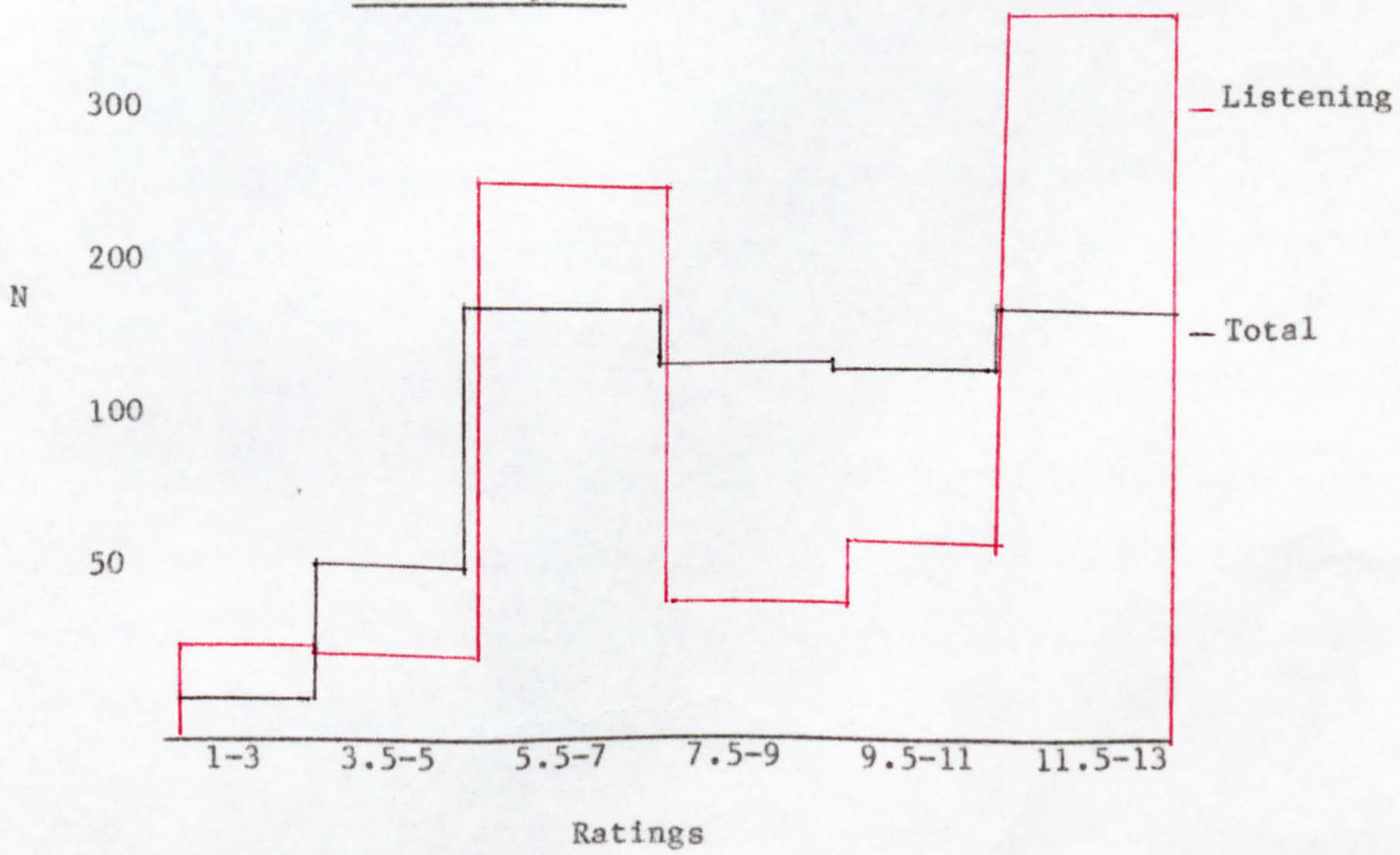


Diagram 5.22 Distributions for tutors' ratings of listening and total compared



5.4.2.4 Consistent differences were noted for the ratings for the two sub-samples. In all skills, ratings for the EPTB sub-sample were lower than for the BCSA sub-sample. The difference in means for Reading was only .3, but for the remaining skills was .9. Distributions are given in Table 5.23.

Table 5.23 Summary of language skills ratings by sub-sample

Ratings	Listening		Speaking		Reading		Writing	
Scale points	EPTB	BCSA	EPTB	BCSA	EPTB	BCSA	EPTB	BCSA
1-3	15	11	24	18	5	7	34	22
3.5-5	14	10	29	29	7	6	28	25
5.5-7	138	112	178	134	114	103	173	143
7.5-9	23	16	18	15	25	14	14	22
9.5-11	31	25	20	29	31	27	26	33
11.5-13	151	193	102	141	185	208	82	114
Total N	372	367	371	366	367	365	357	359
Mean rating	8.9	9.6	7.8	8.6	9.6	9.9	7.5	8.3
SD	3.0	2.9	3.0	3.1	2.7	2.8	3.0	3.1

The standard deviations for both samples were almost identical, although the BCSA sub-sample deviation was 0.1 greater for all skills except writing. The range of ratings was wide for both samples with the BCSA sub-sample having a smaller number of cases from 1 to 5.5 in all skills except writing and a greater number from 11.5 to 13 in all skills.

5.4.3 Tutors were asked to evaluate their students' improvement in English since the previous October, ie the beginning of the academic year. Three options were given: considerable improvement, some improvement, or no improvement.

No responses were given for 20% of the sample, but of the rest more than 95% were adjudged by their tutors to have made improvement, almost a third having made considerable improvement. Responses are summarised in Table 5.24 below.

Table 5.24 Improvement in English noted by tutors

	Improvement	N	%
1.	Considerable	219	32.9
2.	Some	416	62.6
3.	None	30	4.5
	Total	665	100.0
	Missing	168	20.2% of total sample

Distribution was almost identical in both sub-samples. It would seem reasonable to conclude that the mere fact of studying through the medium of English has a positive effect on the proficiency of a student over a period of up to 6 months.

5.4.4 In question 4 tutors were simply asked whether the student had received any tuition in English since the previous October - or the beginning of the course. By this was intended concurrent tuition in English once the main academic session had started - as distinct from pre-sessional tuition arranged by the British Council.

No responses were given for 20% of the sample, and tutors replied that in a further 17.6% of the cases they did not know. Only 181 (21.7% of the sample) were known by their tutors to have received tuition in English since the beginning of the academic year or course. Responses are summarised in Table 5.25 below. Again there was no substantial difference between the EPTB and BCSA sub-samples' distributions, the only response varying by more than 2% from the total being that for "NO".

Table 5.25 In-session tuition in English

Response	N	%
Yes	181	21.7
No	337	40.5
Don't know	147	17.6
Missing	168	20.2
	833	100.0

5.4.5 Tutors were asked to write the qualification being aimed at by the student at question 5. This question was inserted to supplement and confirm the information obtained from the British Council and incorporated in the information already reported (see 5.3.5.1 and 5.3.5.2 above).

5.4.6 Question 6 gave tutors an opportunity to make further comments on the reverse of the rating form. Over 60% of the tutors made at least a one sentence comment, and in most cases much more. These comments tended to convey the following types of information:

- further explanation of responses to one or more of the questions
- further comments on the student in question
- general comments about standards of English of overseas students
- comments on the British Council's handling of students from overseas
- comments on the questionnaire.

Comments on 10% of the sample were analysed closely in the light of these factors. It was found that just over 80% of the comments related to further explanation of the responses given or to further information on the individual student's ability. Improvement in English was also often discussed. Further background on the type of course being followed was sometimes given.

The distinction between the first two kinds of response was not always clear. Examples of such comments are:

> (Further explanation of responses) 094 (Indonesia)
>
> I had difficulty in filling in 1A and B. Mr S's general ability in English is just adequate. This arises because he

has some difficulties in understanding the spoken Engish but if he has time he does understand perfectly: occasionally he says he has understood but one subsequently realises that he has not understood completely. He does not, in fact, make an undue number of mistakes in his spoken or written English.

(Further comments on the student) 112 (Yemen)

1. Is currently attending a whole time Internal Medicine Course
2. Scored 48% on his ELBA test on 17 January
3. Saw me on 8 February: felt the medicine course too much of a strain on his English comprehension
4. I suggested he should attend what lectures he could and recommended he take an English course arranged for Medical Postgraduate by .....
5. EPTB - 32.1 (April)

This comment is clearly one of the more complete comments but represents the different kind of information and comment often given by tutors.

About 7% of the comments analysed alluded to questions of shyness, anxiety and motivation. One tutor even suggested that more extrovert personalities stood a much stronger chance of being adequate in their English. Tutors who made these comments found it difficult to determine whether shyness hampered students in the development of their English or whether it simply concealed some of their ability.

Some tutors made comments of a general nature which their own student's problems highlighted. The following comment is perceptive and represents one of the more comprehensive comments made.

148 (Burma)

Two points seem worthy of mention:
1. It was appreciated, in Mr S's case, that because his first degree was not taken in entirely appropriate sciences, he might have more than usual difficulty in settling into the course. Even graduate chemists have to adapt to the jargon of polymer chemistry and polymer physics - and Mr S is a graduate agriculturalist! However, he has adapted remarkably well.

2. A weakness in spoken English is manifest with almost all Asiatic students. They speak so quickly, and run words into one another to an even greater extent than do English nationals; they also tend to start a question or comment part way through the subject rather than at a logical beginning. All this must work against them, on a study-course, if only because the lecturer (being unable to fully comprehend their question or point) must often pass on without giving a satisfactory answer.

Other tutors commented on certain nationalities staying together and thus not using their English very much, or else commented on the fact that students accompanied by their families often lacked sufficient opportunity to practise their English.

5% of the comments analysed related to the questionnaire itself. These were sometimes mildly critical. The most critical comment was:

1. Difficult to comment because your form doesn't distinguish betwen English and American, eg 'native ability' of which country? .....
4. You ought to take courses in Social Psychology to learn something about questionnaire construction.

Eight of the eighty-three responses analysed referred to question 1A and noted or implied that the gap between the categories 'just adequate' and 'completely adequate' was too great. It was felt that some students were rather better than 'just' adequate but not 'completely' adequate. This point was accepted by the investigator and question A was expanded to four choices for the second investigation. Question 1B was intended to overcome this difficulty by offering more categories, but it became clear that the very first question facing the assessors needed to be acceptable to them and formulated more satisfactorily.

5.4.7 Conclusions on the tutors' ratings

The English Ability Rating Form was completed and returned by a very satisfactory proportion of tutors. The investigator was struck by the fact that no tutor seriously questioned the utility and validity of the rating instrument. It could have been that among those who did not return the questionnaire a large number did not recognise its validity, but not one tutor wrote to say as much.

Respondents seemed tc have had no difficulty deciding on the adequacy or inadequacy of their students' English and the nature of the responses indicated that tutors responded seriously and to the best of their ability. The differentiation between abilities in different skills (question 2) was particularly striking.

The ratings seemed to possess a high degree of validity and produced the information the investigaton required both to proceed with further analysis to determine the validity of EPTB and BCSA and to modify the instrument for the second investigation.

5.5 Conclusions on the First Investigation

One of the main purposes of the first investigation - the trialling of the English Ability Rating instrument - was achieved. Background data had been collected satisfactorily and a large enough sample obtained. Cooperation of the British Council offices affected and of the tutors was greater than the investigator had anticipated, and logistical problems were overcome. The rating instrument appeared to have an acceptable degree of reliability at .85. Feedback from the tutors resulted in only one minor change to the format and content of question 1.

The second purpose of the first investigation was to prepare a framework for analysis of the data. This is discussed in the next chapter.

CHAPTER 6 DISCUSSION OF RESULTS OF THE FIRST INVESTIGATION

6.0 Introduction

The major task following the collection of the data was to reach some preliminary conclusions on the relationship between the tutors' ratings and the pre-departure assessments for both the EPTB and BCSA samples (see paras 2.9.3 and 2.9.4 above). In addition, the relationship had to be re-examined in the light of certain intervening variables (see 2.9.5 above). The conclusions reached were preliminary because of the absence of the communicative proficiency measure, but the methods of analysis were modified for use in the main investigation a year later.

6.1 The Method

Two methods of comparison were employed - correlation and contingency tables. The first method is a mathematically precise method of comparing scores on pairs of measures. The second is a simpler procedure dealing with actual and expected frequencies of cases attaining certain criteria or scores.

6.1.1

Three assumptions are normally made for the use of the Pearson product moment coefficient of correlation. The scores have been obtained in independent pairs unconnected with each other; the two variables correlated are continuous; and there is a rectilinear relationship between the two variables (Guilford, 1973: 94-95). The third assumption is the most important.

In the present investigation it was felt that these assumptions were valid. Pre-departure scores and tutors' ratings were totally unconnected, both were continuous variables, and the basic

assumption was that the relationship between the two would be rectilinear - that is it was anticipated that those students who had obtained low or high scores on BCSA or EPTB would most likely obtain correspondingly low or high ratings from their tutors. There was the recognition of the possibility of the intervention of other variables that might affect the rectilinearity of the relationship, but correlation was nevertheless considered to be a valid method of comparison.

6.1.2 The use of contingency tables was decided in order to give another dimension to the results and because the product moment correlations were not expected to be high. This latter expectation arose mainly because of intervening variables of contact with English, such as the following of courses of remedial English and the fact of having to use English every day. The presence or absence of courses in remedial English had been controlled for in the collection of background data, but it was not possible to control for the effect of contact with English as no subjects were pursuing their studies outside an English-speaking environment. Tutors' responses to the question about improvement in English had shown that working in English had an obvious effect on students' proficiency (see para 5.4.3 above), and many of the 4.5% who were deemed to have made no improvement were so proficient on arrival that there was little room or need for further improvement.

Contingency tables established the extent of the accuracy of the predictions by means of 2 x 2 tables or by means of tables accounting for a greater number of categories, eg 3 x 3 tables.

In addition chi-squares were computed to determine firstly whether there was a relationship between the variables on the different axes and secondly to determine whether or not the relationships could be attributed to chance. Significance beyond the .05 level was accepted as evidence of relationships being independent of chance factors.

6.2 Comparison by Correlation

Totals representing the EPTB Part 1 scores and the 'notional' overall grade for BCSA were correlated with tutors ratings on the 6 point scale at question 1B of the rating form and with the notional proficiency rating total, obtained by summing ratings given in answer to question 2 of the rating form.

Because there were two different pre-departure assessments of English used there was no measure for the total sample that could be correlated with tutors' ratings. However, since approximately 40% of the EPTB sub-sample were given supplementary oral and writing assessments it was possible to identify a third sub-sample consisting of the BCSA sub-sample and 40% of the EPTB sub-sample. Their performance on subjective oral and writing tasks could be compared with tutors' ratings of their abilities in spoken and written English and with the notional tutors' ratings. Three sets of correlations were therefore obtained and are reported separately below.

6.2.1 Table 6.1 summarises the correlations obtained where scores on EPTB Pt 1 were compared with tutors' ratings as expressed in response to question 1B of the rating form (R1B) and with their

skills ratings when totalled in question 2(RT). N indicates the number of cases and S the significance level.

Table 6.1 Correlations: EPTB Pt 1 with tutors' ratings

(a) EPTB Pt 1 with tutors' ratings (R1B)	.319 N = 371 S = .001
(b) EPTB Pt 1 with tutors' ratings (RT)	.319 N = 354 S = .001

Both correlations were identical and both highly significant. Further correlations of sub-test scores with tutors' ratings were also obtained and reported in Table 6.2

Table 6.2 Correlations: EPTB sub-tests with tutors' ratings (RT)

Sub-test	r	N	S
1. Phonemic discrimination	.06	319	NS
2. Intonation	.21	319	.001
3. Reading comprehension	.21	319	.001
4. Grammar	.26	319	.001
5. Reading speed	.24	154	.001
6. Oral	.22	136	.005
7. Writing	.10	138	NS

It was noted that all the further correlations were positive but not as high as the correlation with EPTB total score. Correlations with sub-tests of grammar and reading speed at .26 and .24 were the highest while the correlations with the tests of phonemic discrimination and writing (subjective) were not significant.

6.2.1.1 About 40% of the EPTB sub-sample had been given pre-departure subjective assessments in their ability to write and to speak English. These assessments were correlated with their British

tutors' ratings of their writing and speaking abilities. The correlations are presented in Table 6.3 below.

Table 6.3 Correlations: EPTB writing and speaking with tutors' ratings

(a) Writing (subjective) with tutors' rating of writing (RW)	.21 N = 138 S = .007
(b) Speaking (subjective) with tutors' rating of speaking (RS)	.17 N = 143 S = .021

Although both correlations were significant at the 5% level they were lower than the significant correlations obtained when comparing sub-test performance with total tutors' ratings (see Table 6.2 above).

6.2.2 Table 6.4 summarises the correlation obtained when comparing the notional total grades on BCSA with tutors' ratings at question 2 on the rating form. No correlation was available with ratings at question 1B.

Table 6.4 Correlations: BCSA total with tutors' ratings

(a) BCSA notional total with tutors' ratings (R1B)	N/A
(b) BCSA notional total with tutors' ratings (RT)	.352 N = 339 S = .001

The correlation was highly significant and higher than the comparable correlation obtained with the EPTB sample. Since pre-departure assessments had been made skill by skill, these were correlated with the tutors' individual skills ratings, as expressed

in question 2 of the rating form, and also with the skills total. All were found to be highly significant.

Table 6.5 Correlations: BCSA skills with tutors' ratings (skills and total)

		N	r		N	r
(a) Listening with listening	(RL)	344	.231	with RT	335	.212
(b) Speaking with speaking	(RS)	344	.316	with RT	336	.295
(c) Reading with reading	(RR)	338	.251	with RT	331	.323
(d) Writing with writing	(RW)	333	.285	with RT	331	.325

Three points emerged from these correlations. Firstly, the correlations with reading and writing and tutors' total ratings were very close to the total correlations reported in Table 6.4 and were just a little higher than EPTB Part 1 scores correlated with total ratings. All the correlations reported for BCSA were higher than those reported for EPTB, with the exception of the .21 correlation of BCSA listening with RT. Thirdly the writing and speaking correlations reported for BCSA were significantly higher than those reported for EPTB in Table 6.3, although in both cases the assessments for writing correlated more highly than the assessments for speaking. The fact that the BSCA sample was much larger may in part account for the latter observation, while the fact that the BCSA and tutors' ratings scales were both arrived at subjectively and were similar in nature may account for the higher overall correlation.

6.2.3 Correlations of pre-departure speaking and writing with tutors' ratings of speaking and writing abilities (RS and RW) for the whole sample are given in Table 6.6.

Table 6.6 Correlations: Speaking and writing with tutors' ratings

		r	N	S
(a)	Speaking with tutors' ratings (RS)	.282	485	.001
	Speaking with tutors' ratings (RT)	.282	470	.001
(b)	Writing with tutors' ratings (RW)	.287	470	.001
	Writing with tutors' ratings (RT)	.283	468	.001

There were no significant differences between the skill with skill correlations and the skill with total correlations, and all were between .28 and .29. The correlations were a little lower than those reported for total skills with total proficiency ratings. This was not surprising as the individual skills formed part of the totals in both cases.

6.2.4 A total of 20 correlations of pre-departure measures with different tutors' ratings have been reported above. The most significant for the investigation were the correlations of totals of pre-departure assessments with totals of tutors' ratings. All were highly significant and similar in value, with two at .319 and one at .352. Skills with skills correlations were also highly significant, except in one case (EPTB speaking with tutors' rating of speaking), ranging from .17 to .316. Nine correlations between sub-tests or skills and tutors' total ratings were obtained. Two of these were not significant at the 5% level and the remainder ranged from .21 to .28 only.

Two features of these correlations stood out. Firstly all were positive, and secondly the range was limited, with 14 of the 20 correlations between .25 and .35.

6.2.4.1 The most important question to be answered at this stage was - do these correlations indicate a positive and significant relationship between performance on EPTB or ECSA and current performance as rated by the tutors? Statistically, the answer to both was 'yes'. The main problem was whether correlations of .32 and .35 amounted to strong enough relationships between the two sets of measures so as to confer some measure of predictive validity to both EPTB and ECSA. A closer consideration of the nature of a product moment correlation coefficient was undertaken.

The establishing of a correlation between two variables assumes ideal conditions, particularly in respect of the population, the continuous nature of the scales, the linearity of the relationships, and freedom from interference from intervening variables. Most experiments involving subjects in real-life situations are not conducted in ideal circumstances, and the present investigation was no exception.

Guilford (op cit: 315) has noted that the coefficient of validity in a restricted group is almost invariably smaller than it would be in an unrestricted group. In these terms the sample in this investigation represented a very restricted group. If all the students who had applied for awards to study in Britain had been accepted and brought to Britain and sent on courses of their choice, the group would have been much less restricted. It would be reasonable to suppose that had that been the case, the sample would have been from five to ten times larger and the English proficiency of the subjects would have been inadequate from the start in at least 50% of the cases. In those circumstances, a closer relation-

ship between pre-departure assessments and tutors' ratings would have been expected. The sample examined in the first investigation, therefore, represented a very restricted group - particularly in respect of under-achievers.

The scales used for EPTB, BCSA and tutors' ratings were continuous, but the use made of the scales by assessors in BCSA and in the tutors' ratings was not continuous. Examination of the distributions of the grades for BCSA presented in Tables 5.11 and 5.12 above show that, with the exception of assessment of writing ability, 85-90% of all grades awarded were in the upper half of the scale with the mode often at the highest point. In the tutors' ratings, it would appear that tutors tended to use particular sections of the scale - eg 11 to 13, and 5 to 7.

Distributions of ratings presented in Table 5.19 and diagrams5.21 and 5.22 above showed the tendency to bimodality and to the use of the upper points of the scale. The resulting skewness of the distributions reduced the extent of the correlation.

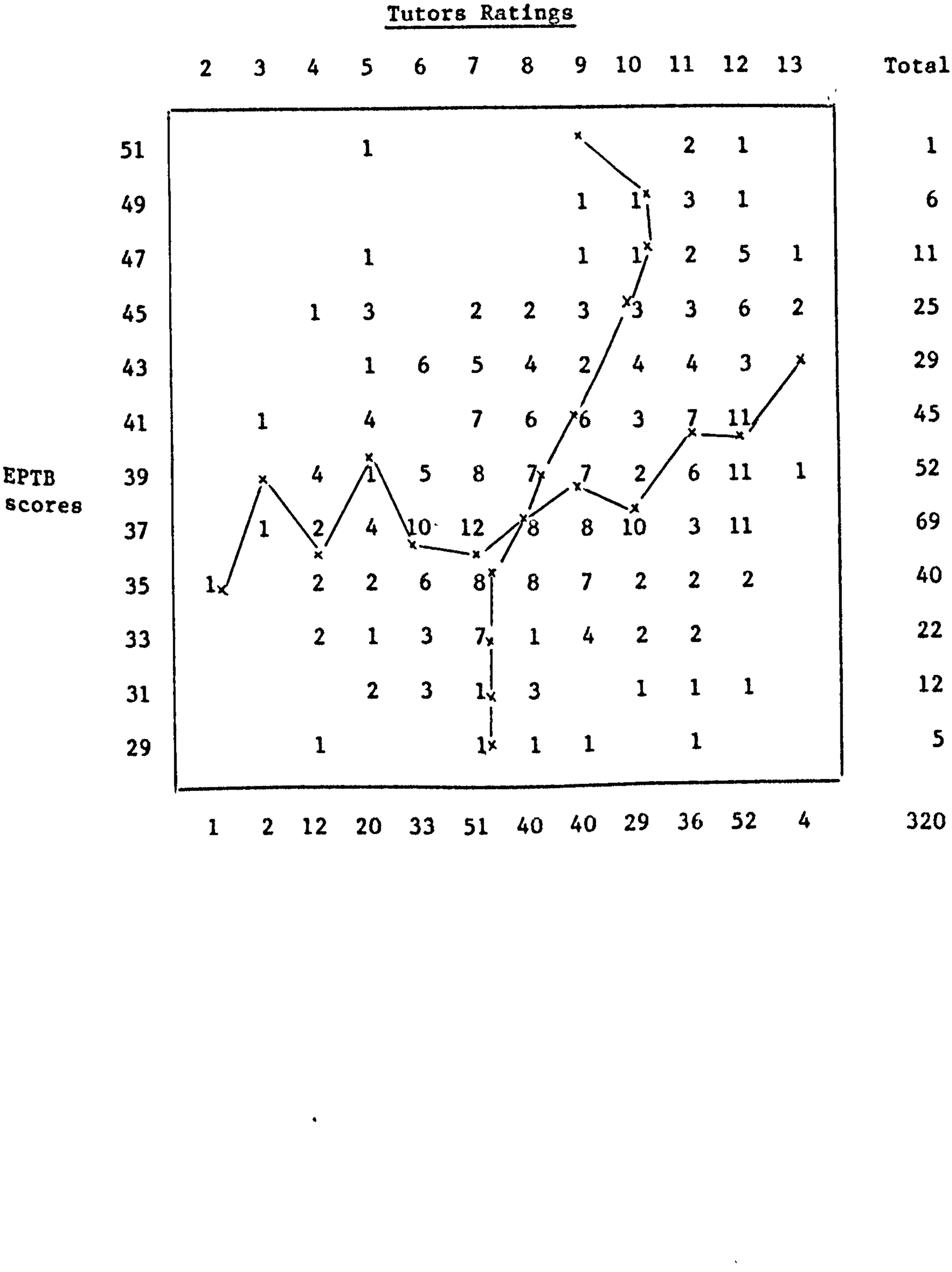
Another feature of the tutors' rating scale was that the upper half of the scale could be considered optional in the sense that the points on the scale represented greater degrees of adequacy. For the purpose of studying in the medium of English, both students and tutors knew that complete adequacy was not essential - though highly desirable. Therefore, it could be argued that although the scale is continuous, neither the student nor the tutor was necessarily looking for performance which exceeded 7 or 8 on the scale. In the case of the student it could have meant that once

he or she felt that that point or threshold had been reached, no more of his or her energies would be directed to language improvement. This might account for the high number of ratings around point 7 on the rating scale.

One of the results of the bimodality and skewness of the scales was to disappoint the expectation of a linear relationship between assessments on the two scales. The first 320 completed cases of the EPTB sub-sample were taken and a scatter diagram produced (see Diagram 6.7). Two regression lines were drawn through the points representing the means of each of the columns (EPTB scores) and each of the rows (tutors' ratings). This showed graphically that the regression lines were curved and irregular in shape as well. It was only in the upper sections of the scales, from columns 7 to 13 and rows 43 to 49, that the lines began to advance in a manner that matched the anticipated regression. This lack of linearity also helped to reduce the size of the correlation.

Finally, in an ideal situation, all other factors should have remained constant so that the same conditions obtaining at the pre-departure assessments should have obtained at the time of the tutors' ratings. Thus students should not have followed any further English classes, they should not have changed their attitudes or motivation to English, they should have studied with no less but no greater effort and so on. Clearly in a situation where the interval between the assessments may have varied from 18 to 8 months, the abilities observed by the tutors had been seriously 'contaminated'. In a 'contamination free' environment the correlation could well have been close to 1.0. But in the situation

Diagram 6.7 Scatter diagram: EPTB and tutors' ratings



that obtained in the current investigation it was clear that conditions had been affected by a quantity of other factors and so only a much smaller correlation could be expected. Guilford has summed this up neatly (op cit: 92):

> 'Thus, in nature, correlations of zero or 1.00 are expected to be the rule between variables when their effects are experimentally completely isolated. The fact that we obtain anything else is due to the inextricable interplay of variables that we cannot measure in isolation.
>
> The practical conclusion to be drawn from this is that a correlation is always relative to the situation under which it is obtained, and its size does not represent any absolute natural fact.'

The situation under which any investigation of predictive validity is carried out must inevitably involve considerable time lapses and consequent exposure to the influence of other variables. In this study the most that it seemed reasonable to expect were positive and significant correlations of pre-departure assessments with subsequent appropriate measures. Correlations of .32 and .35 obtained with tutors' ratings, supported by almost 70 more positive and significant correlations from 47 sub-samples discussed later in this chapter, suggested a strong and very positive relationship between the two measures. This relationship was further tested by establishment of contingency tables which are discussed in the next section.

6.3 Comparison by Contingency Tables

On the basis of the scores obtained, the subjects were divided into categories for the various measures. The number of cases were established for each of the categories and then tabulated. The simplest tables were 2 x 2 tables where the categories were inadequate and adequate. More complex tables using a greater number of categories

were also plotted. In all cases chi-squares were computed. Separate tables had to be established for both the EPTB and the BCSA sub-samples.

6.3.1 The EPTB sub-sample was grouped into three according to scores on the Part 1 total. Group 1 consisted of those who had obtained scores of 33.9 or less and whose English was considered inadequate (see Table 5.9 above). Group 2 consisted of those who had scored from 34.0 to 39.9 and whose English had been considered just adequate and who had been recommended to follow remedial English classes. They numbered 51% of the sample. Group 3 consisted of those who had scored 40.0 and above and whose English was considered adequate for study in Britain. 37% of the sample were in this group. The groups were then further subdivided according to the tutors' ratings expressed in question 1A of the English Ability Rating form (R1A). Members of group 1 were considered to have inadequate English for their studies, those in group 2 to possess English that was just adequate and those in group 3 to posses completely adequate English. The breakdown obtained, or cross tabulation, is given in Table 6.8 below:

Table 6.8 Cross tabulation: EPTB with tutors' ratings (R1A)

		Ratings (R1A) 1	2	3	Row total
EPTB	1	1	30	11	42
	2	9	106	75	190
	3	2	47	89	138
Column total		12	183	175	370

6.3.1.1 This cross tabulation provided the basis for a number of contingency tables. The first table, a two way table, looked at the dichotomous prediction of the EPTB scores:

(i) English proficiency inadequate, or
(ii) English proficiency adequate

and then at the two basic assessments of the tutors,

(iii) English is inadequate, or
(iv) English is adequate.

To obtain (i) members of EPTB group 1 (row 1 in Table 6.8) were taken. To obtain (ii) cases in EPTB groups 2 and 3 (rows 2 and 3 in Table 6.8) were merged. For (iii) cases in R1A group 1 (column 1 in Table 6.8) were taken, and R1A groups 2 and 3 were merged to correspond to (iv).

The expectation was that all cases at (i) - ie with inadequate English at the outset - should reappear in the category (iii) - ie with inadequate English now. Similarly all cases at (ii) were expected to appear in category (iv) - ie adequate English at the outset and now. Cases appearing in those categories were considered 'hits', and cases in rows (i) and (ii) who were assigned to (iv) and (iii) respectively by the tutors - ie inadequate students whose English had become adequate and students with initially adequate English whose English was subsequently rated inadequate - were considered 'misses'.

Table 6.9 Contingency table: EPTB with tutors' ratings (R1A)

(no of cases)

		R1A (iii)	R1A (iv)
EPTB (i)	(i)	1*	41
	(ii)	11	317*

(percentages)

		R1A (iii)	R1A (iv)
EPTB (i)	(i)	.3*	11.1
	(ii)	2.9	85.7*

In the table asterisks have been inserted in the boxes where expectations were met (hits). The boxes without asterisks represent misses. On the basis of the simple dichotomies the table indicates a rate of hits of 86%. This was considered highly satisfactory, especially since the misses were made up largely of cases whose English was originally thought to be inadequate but was subsequently deemed at least just adequate. They represented cases of error of prediction which nevertheless brought them benefit. The 2.9% of cases, however, who were assigned to column (iii) and row (ii) represented cases where error in prediction had worked against them. They were the real casualties. It says much for the accuracy of the overall prediction that this proportion was limited to less than 3% of the total sub-sample.

In case R1A should have proved an unsatisfactory measure for comparison, further tables were worked out using questions 1B and questions 2 of the English Ability Rating form. However, in view of the high intercorrelations obtained between responses to questions 1A, 1B and skills total, it was anticipated that the make up of the tables would be similar to the one already reported. Category (iii) was taken to include those cases whose English was considered to be either well below satisfactory and to display considerable deficiencies (see Table 5.16 above), or to include those with notional proficiency ratings of 4 or less (see diagram 5.18 above). Category (iv) was taken to include those whose English was considered just adequate and better, or those with notional proficiency ratings of 4.5 or above. The cut off at 4 inclusive was chosen because it represented the lower third of the scale - in line with

the 3 and 6 point scales of questions 1A and 1B - and also represented a point half way between the first letter of the word inadequate and the final letter of the word adequate as they appeared above the scale on the rating form (see para 5.4.2 above).

Tables 6.10 and 6.11 present the contingency tables for EPTB with rating form question 1B (R1B) and question 2 total (RT) respectively.

Table 6.10 Contingency table: EPTB with tutors' ratings (R1B)

(no of cases)

EPTB		R1B (iii)	R1B (iv)
	(i)	1*	41
	(ii)	13	316*

(percentages)

EPTB		R1B (iii)	R1B (iv)
	(i)	.3*	11.0
	(ii)	3.5	85.2*

Table 6.11 Contingency table: EPTB with tutors' ratings (RT)

(no of cases)

EPTB		RT (iii)	RT (iv)
	(i)	3*	38
	(ii)	13	300*

(percentages)

EPTB		RT (iii)	RT (iv)
	(i)	.8*	10.7
	(ii)	3.7	84.7*

In the above tables the percentage of hits is identical at 85.5%, differing by only 0.5% from the percentage of hits obtained using question 1 (R1A). There were very minimal variations in distributions using these two measures, although the number of detrimental misses increased slightly.

6.3.1.2 These tables could be analysed from another standpoint, that of the increase in efficiency over chance. If there were no

information in the form of tutors' ratings available and if no predictive criterion data had been made available by Davies when he was developing EPTB, and if candidates were still divided into those with less than 34 and those with more than 34, the predictive power of the scores would be no better than that attributable to chance. Thus we could not reasonably expect more than half the candidates with scores of less than 34 to prove inadequate in their English proficiency, nor could we expect more than half the candidates with 34 or more to prove adequate. By making use of the tutors' ratings and similar criterion measures, and assuming their validity, prediction from EPTB or BCSA should become more efficient.

Table 6.12 shows a contingency table with the same groupings at (i) and (ii) but with chance groupings at (iii) and (iv). Expected totals (hits) are marked as usual with an asterisk.

Table 6.12 Contingency table: EPTB with chance

(percentages)

		chance (iii)	chance (iv)
EPTB (i)		6*	6
	(ii)	44	44*
	Column total	50	50

The proportion of hits with chance can only be 50%. The tables with the tutors' ratings, with the proportion of expected outcomes at between 85% and 86% are therefore much more efficient. They have predicted 36% cases more accurately representing an increase of 36/50% increase in efficiency, which is equivalent to a 72% increase in efficiency over chance. This offers further evidence

of the basic predictive power of EPTB and of the strength of the relationship between EPTB and the tutors' ratings.

6.3.1.3 Further contingency tables were worked out using three categories per axis as in the cross tabulation at Table 6.8 above. The frequencies in that table are expressed as percentages in Table 6.13 below.

Table 6.13 Contingency table: EPTB with tutors' ratings (R1A) - %

(percentages)		Ratings - R1A 1	2	3	Row total
EPTB	1.	0.3*	8.1	3.0	11.4
	2.	2.4	28.6*	20.3	51.3
	3.	0.5	12.7	24.1*	37.3
Column total		3.2	49.5	47.3	100.0 (N = 370)

Raw chi-square = 30.07 df 4 S = 0.000

It was expected that the cases with inadequate scores on EPTB (row 1) would also be deemed inadequate by their tutors (column 1). This was not borne out by tutors' ratings since most were deemed just adequate (column 2). It seemed reasonable to expect that EPTB borderline cases (in row 2) would prove borderline, or just adequate, when rated by their tutors (column 2). Well over half that group were found to be so. Finally it was expected that those with adequate scores on EPTB would be found completely adequate by their tutors. This was confirmed in almost two thirds of the cases in this group (row 3, column 3).

Exactly 53% of the cases were found in those cells marked with an asterisk. Since only 33% of the cases could be expected in these cells solely as a result of chance, this represented an efficiency of 60% greater than chance. The chi-square was calculated at 30.07 with 4 degrees of freedom and was well beyond the 1% level of significance. This table pointed to the strength of the relationship between the two sets of measures.

6.3.1.4 A contingency table using three categories for tutors' skill ratings totals (RT) was also produced. RT categories were:

1 = cases with total ratings of 4 or below;

2 = cases with ratings of 4.5 to 6 inclusive; and

3 = cases with ratings of 6.5 or more.

It was very difficult to decide on the dividing line between categories 2 and 3, just adequate and more than just adequate. Since category 2 represented a borderline category, the ratings band could not be too extensive. Moreover the rating form was printed in such a way that the printed word <u>adequate</u> extended from point 5 to point 7 on the scale. It was felt that the category of just adequate could hardly extend beyond the word <u>adequate</u>, and so points 5.0, 5.5 and 6.0 were treated as borderline cases and points 6.5 and 7.0 were treated as points representing fully adequate proficiency. This did not correspond with the way the tutors had actually rated. There was a very definite peaking of ratings at points 6.0, 6.5 and 7.0 on the scale. It was inferred that these represented a judgement to the effect that students' English was generally adequate but that there was plenty of scope for a higher standard and greater proficiency. In retrospect it

would have been preferable to have extended category 2 to include point 7 on the scale. The less satisfactory grouping provided some useful information, however, as the distributions were different. The table proved less efficient than the table with R1A since only 44% of the cases were hits.

Table 6.14 Contingency table: EPTB with tutors' ratings (RT) - 9 cells

(percentages)		RT 1	RT 2	RT 3	Row total
EPTB	1.	0.8*	2.5	8.2	11.6
	2.	3.1	10.7*	37.0	50.8
	3.	0.6	4.8	32.2*	37.6
Column total		4.5	18.1	77.4	100.0 (N = 354)

The greatest frequencies were consistently found in column 3. Although this produced a very accurate rate of expectancy for row 3 and column 3, 77% was probably too great a proportion of subjects to be rated completely adequate. The proportion rated in that category in question 1 had been 53% (see para 5.4.1 above) and since correlations of 1A and 1B with question 2 had been well over 0.8, the distribution for column 3 in the table under discussion should have been nearer 53%. The attribution of categories to the total skills rating was therefore considered unsatisfactory and the relevance of the above table was weakened. This is supported by the chi-square which was only just at the 5% level of significance.

6.3.2 The BCSA sub-sample was grouped into three categories on the basis of the notional 'total' grades obtained on their pre-departure

assessment (see Table 5.11 above). Group 1 consisted of subjects with an overall grade of C+ or less. Although the definitions for grade C performance suggested that students with this level of English should be able to cope with their studies after remedial English, it was British Council Headquarters practice to consider such candidates as inadequate. Group 2 comprised those students with an overall grade of B and whose English was considered likely to be just adequate. Students in that category were normally given remedial English before beginning their studies. Group 3 comprised those students with an overall A or B+ and who were considered to be entirely adequate in their English proficiency. They accounted for approximately two-thirds of the sample for whom complete data was available. Groupings according to tutors' ratings were effected using the same criteria as for the EPTB sample, ie following the categories in question 1A of the rating form (R1A) as in para 6.3.1 above. The cross tabulation thus obtained is given in Table 6.15 below.

Table 6.15 Cross tabulation: BCSA with tutors' ratings (R1A)

		Ratings (R1A) 1	2	3	Row total
BCSA	1	4	22	23	49
	2	7	38	38	83
	3	7	67	158	232
Column total		18	127	219	364

6.3.2.1 A 2 x 2 contingency table was drawn up on the basis of the above data. The expectation was that members of BCSA group 1 would

also be deemed inadequate by their tutors, Ratings group 1, and that members of BCSA groups 2 or 3 would be deemed adequate. In the 2 x 2 tables that follow the categorisations of the row cells and the column cells are the same as for the EPTB sub-sample and are spelled out in para 6.3.1.1 above. Table 6.16 contains the contingency tables expressed first in numbers of cases and secondly in percentages.

Table 6.16 Contingency table: BCSA with tutors' ratings (R1A)

(no of cases)

		R1A (iii)	R1A (iv)
BCSA	(i)	4*	45
	(ii)	14	301*

(percentages)

		R1A (iii)	R1A (iv)
	(i)	1.1*	12.3
	(ii)	3.8	82.8*

The rate of hits was 83.9%, just 2% lower than that obtained for EPTB. Nevertheless, it was considered a very satisfactory rate. The rate of misses was slightly higher than for EPTB. The 3.8% placed in row (ii) and column (iii) were students whose English proficiency was predicted as likely to be satisfactory but whose tutors rated their English as inadequate.

Tables using the responses at questions 1B and 2 of the English ability rating form were also prepared. The establishing of categories (iii) and (iv) for the ratings at 1B (R1B) and for the skills total (RT) were the same as for the EPTB sample, as outlined in para 6.3.1.1 above.

Tables 6.17 and 6.18 give the contingency tables for BCSA with R1B and RT respectively.

Table 6.17 Contingency table: BCSA with tutors' ratings (R1B)

(no of cases)		R1B (iii)	R1B (iv)
BCSA	(i)	4*	45
	(ii)	16	300*

(percentages)		R1B (iii)	R1B (iv)
BCSA	(i)	1.1*	12.4
	(ii)	4.5	82.0*

Table 6.18 Contingency table: BCSA with tutors' ratings (RT)

(no of cases)		RT (iii)	RT (iv)
BCSA	(i)	2*	47
	(ii)	10	298*

(percentages)		RT (iii)	RT (iv)
BCSA	(i)	0.6*	13
	(ii)	2.8	83.6*

There were no significant variations in the three tables. The number of hits varied only from 83.1% to 84.2% depending on the rating criterion used. The detrimental misses varied a little more from 2.8% to 4.5% of the cases but in all cases the numbers were small.

6.3.2.2 As with the tables relating to the EPTB sub-sample, the tables proved much more efficient than expectations based on chance. Increases in efficiency were 66.2%, 67.8% and 68.4%, marginally lower than those for the EPTB sub-sample.

6.3.2.3 Contingency tables using three categories per axis were also established. The three categories on the BCSA axis corresponded to the groupings outlined in 6.3.2 above, while the groupings for R1A corresponded to the three responses given in question 1A on the rating form. The numbers of cases were those displayed in the cross tabulation in Table 6.15 above, and are expressed in percentages in Table 6.19 below.

Table 6.19 Contingency table: BCSA with tutors' ratings (R1A) - 9 cells

(percentages)		Ratings (R1A) 1	2	3	Row total
	1.	1.1*	6.0	6.3	13.5
BCSA	2.	1.9	10.4*	10.4	22.8
	3.	1.9	18.4	43.4*	63.7
Column total		4.9	34.9	60.2	100.0 (N = 364)

Raw chi-square = 18.15 df 4 S = 0.001

The proportion of hits obtained was 54.9%, almost 2% higher than in the equivalent table for the EPTB sub-sample (see Table 6.13 above). BCSA predictions at the B+ and A level were the most efficient (row 3 and column 3) and accounted for 80% of the hits. This was in part due to the skewed distributions of the BCSA sample both along the BCSA axis and the ratings axis.

When a contingency table was established using the categories for question 1B on the ratings form the distributions did not vary more than 1% per cell, and the proportion of hits was 53.5%. Chi-squares were computed for the tables with R1A and R1B and were significant beyond the 1% level.

6.3.2.4 A contingency table using the tutors' 'total' skills ratings (RT) was also produced. The categories established were as for the EPTB sample (see para 6.3.1.4 above).

Table 6.20 Contingency table: BCSA with tutors' ratings (RT) - 9 cells

(percentages)		Ratings (RT) 1	2	3	Row total
BCSA	1.	0.6*	2.5	10.6	13.7
	2.	1.7	3.9*	16.8	22.4
	3.	1.1	7.3	55.5*	63.9
Column total		3.4	13.7	82.9	100.0 (N = 357)

Raw chi-square = 9.6 df 4 S = 0.048

The chi-square was just beyond the 5% level of significance and the proportion of hits was exactly 60% - the highest of the 3 x 3 tables. This represented an efficiency of 80% better than chance. The strength of the relationship shown in this table may be in large part due to the distribution of the tutors' ratings as represented by the cut-offs. Over 80% of the sample were adjudged by their tutors to be in the upper half of the scale of adequacy. This figure is much larger than the number of ratings of completely adequate (see R1A column 3, Table 6.19) which accounted for over 60% of the sample. It was therefore to be expected that about 60% of the 80% - or half the BCSA sub-sample - would be placed in the adequate pre-departure and adequate tutors' ratings (row 3 and column 3) cell. On the other hand if the relationship between the assessments had not been strong, the number of hits would have been fewer.

6.3.2.5 The contingency tables, supported by the raw chi-squares obtained, confirmed the evidence of the strong positive relationships between both pre-departure measures and the tutors'

ratings which had already been obtained from the correlations. The 2 x 2 contingency tables showed expectations-confirmed (hits)- at the rate of 85% and 86% for the EPTB sub sample, and 83% and 84% for the BCSA sub sample. Predictions were significantly more efficient than those based on chance by 66% to 72%.

When 3 x 3 tables were established a new element provided further evidence of the strength of the relationships. Although the tutors' ratings categories were identical for the two sub-samples, the distributions for the categories varied from one sub-sample to the other, see para 5.4.11 above. Similarly, although the pre-departure assessment categories used in the tables for the two sub-samples were identical, the measures on which they were based were completely different and the distributions within the categories also differed significantly, as Table 6.21 shows. Whereas the EPTB sub-sample distribution approximated to a negatively skewed normal distribution, the BCSA sub-sample was distributed in a rising curve.

Table 6.21 Distributions of EPTB and BCSA categories compared

Category	% of cases EPTB	BCSA
1. Inadequate English	11.4	13.5
2. Just adequate English	51.3	22.8
3. Adequate English	37.3	63.7

In spite of these differences the rate of hits ranged from 53% to 60% in three of the tables (only 43.7% in the EPTB/RT table) with efficiency at between 60% and 80% better than chance. These results, in the opinion of the investigator, testified to the robustness of the tables and to the predictive power of the test instruments.

6.4 Other Variables

In the foregoing discussion the effects of intervening variables was discussed, especially with reference to the correlations (see para 6.2.4.1 above). The final hypothesis tested in this enquiry (see para 2.9.5 above) was that the relationships between EPTB or BCSA and the criterion measures for groups affected by intervening variables were significant and positive. This therefore necessitated examination of the data with specific reference to each of the relevant variables. The first was the following or omission of a pre-sessional remedial English course. Again the groups had to be divided according to the EPTB and BCSA sub-samples.

6.4.1 Each of the sub-samples was subdivided into a no remedial group and a plus remedial group. Certain expectations were formulated for each of the groups. Firstly, it was anticipated that the groups who had had no remedial English tuition would be less 'contaminated' than the 'plus remedial' group and would consequently display higher correlations between pre-departure and tutors' assessments. It was thought that those correlations could even be higher than for the total EPTB and BCSA sub-samples. It was anticipated that the equivalent correlations for the 'plus remedial' group would be lower and might even be negative in view of the effort that had been made to upgrade these subjects' English. It was further anticipated that the difference between the adequacy of the two groups' English would have been considerably narrowed by the time the tutors carried out their ratings.

6.4.1.1 Membership of the groups was as follows:

EPTB:	No remedial	175	Plus remedial	264
BCSA:	No remedial	179	Plus remedial	215
Total	No remedial	354	Plus remedial	479

EPTB scores for each of the remedial groups were correlated with tutors' ratings at question 1B (R1B) and with the skills 'totals' (RT) at question 2. Similar correlations were computed for the BCSA remedial groups. Correlations were obtained as follows:

EPTB Pt 1 total with ratings (R1B) - No remedial group - r = .28
EPTB Pt 1 total with ratings (RT) - No remedial group - r = .28
BCSA 'totals' with ratings (R1B) - No remedial group - r = .26
BCSA 'totals' with ratings (RT) - No remedial group - r = .28

All the correlations were significant beyond the 1% level.

These correlations were not as high as expected, in all cases being lower than total EPTB Pt 1 or BCSA with tutors' ratings. However, it had to be borne in mind that the samples were very restricted, using Guilford's terms, and the correlations were therefore considered satisfactory. A closer look at the results showed that many subjects who had been deemed adequate on pre-departure tests proved only just adequate according to their tutors. Similar numbers were rated completely adequate by their tutors although pre-departure ratings deemed them only just adequate.

Correlations for the 'plus' remedial group were obtained as follows:

EPTB Pt 1 total with ratings (R1B) - plus remedial - r = .12
EPTB Pt 1 total with ratings (RT) - plus remedial - r = .17
BCSA 'totals' with ratings (R1B) - plus remedial - r - .2
BCSA 'totals' with ratings (RT) - plus remedial - r = .2

Three correlations were significant at the 1% level, with the lowest correlation at the 5% level. As anticipated, these correlations were lower than those for the 'no remedial' group but not as low as zero, and still significant.

6.4.1.2 The correlations for the 'no remedial' groups were not as high as expected but nevertheless were superior to those for the 'plus remedial' groups. The conclusion drawn from these relationships was that the presence of pre-sessional remedial English courses was one variable that affected the overall relationship between the pre-departure and tutors' ratings. For the group that had taken remedial English the relationship was much less significant and positive than for the 'no remedial' group. There was no substantial difference between the sub-samples except that 'contamination' seemed to be slightly greater for the EPTB sub-sample.

One final point was examined - the gap in proficiency between the sub-samples. This was established by comparing the means and standard deviations for the EPTB or BCSA scores and those obtained for the 'total' tutors' ratings. These are presented in Table 6.22 below.

Table 6.22 Means and SDs for the remedial English sub-samples

			Pre-departure	Rating (RT)
EPTB Pt 1	Total sub-sample	M	38.9	35.1
		SD	4.5	10.1
	No remedial	M	42.03	36.8
	Plus remedial	M	36.08	31.5
	Difference	-	over 1 SD	½ SD
BCSA total	Total sub-sample	M	139.4	36.4
		SD	20	10.6
	No remedial	M	152.5	40.0
	Plus remedial	M	128.8	33.5
	Difference	-	over 1 SD	2/3 SD

The table shows the real difference which existed in proficiency at the time of the pre-departure assessments - in both the EPTB and BCSA sub-samples the difference was over 1 standard deviation.

The experience of remedial English reduced the gap between the two groups but did not close it. Although the average length of remedial tuition was almost 8 weeks, this was not enough to enable the weaker students to catch up with the students who had no need of remedial English.

6.4.2 The next variable to be considered was country of origin. By dividing the sample into a number of groups according to country of origin it was thought that it would then be possible to ascertain whether a range of correlations was being obtained and whether certain country groups were producing high correlations. When high correlations were obtained, it was not possible to find the reason, but it could have been as a result of consistency of linguistic and educational background as well as a result of the range of abilities represented by the group. For country groups that had been assessed subjectively, it was thought that a higher correlation might also be an indicator of the consistency of the pre-departure assessment.

The EPTB sub-sample was slightly larger than the BCSA sub-sample but it covered a much more restricted range of countries. Consequently country sub-samples tended to be larger. 15 was chosen as the minimum size for a country sub-sample. However, Afghanistan with 12 was also included because it was the largest BCSA country group from Asia, and Peru was included because it was the only country which used both forms of assessment, the EPTB N being 14 and BCSA N being 10. The danger of having such small sub-samples was that some of the correlations would be non-significant.

6.4.2.1 13 country groups were taken from the EPTB sub-sample and 7 country groups from the BCSA sub-sample. The mean age for each group was recorded as well as the mean EPTB Pt 1 score or mean BCSA total grade. The mean RT rating was also noted. Comparison of these factors shed light on, for example, any sub-group sending older students, or any sub-group that came with particularly high or low pre-departure assessments. For each group two correlations were produced: EPTB Pt 1 score or BCSA total grade (DT) with R1B and RT. The correlations and other supporting data are given in Table 6.23 below.

Table 6.23 Correlations: EPTB and BCSA with R1B and RT by country of origin

Country group	N	Age	DT	RT	r DT/ R1B	r DT/ RT
EPTB sub-sample	439	29.96	38.89	8.45	.319	.319
Mexico	69	27.1	38.75	8.49	.373	.522
Thailand	55	29.9	39.25	7.82	.632	.492
Sudan	41	30.2	39.09	9.25	.376	.306
Indonesia	39	34.9	38.36	7.36	.294	.497
Turkey	35	29.3	41.1	9.27	.674	.628
Iran	23	33.2	38.19	7.23	.179*	-.001*
Egypt	23	32.9	38.22	9.46	.68	.332
Burma	26	31.3	38.2	9.22	-.258*	-.176*
Algeria	21	21.5	41.7	7.99	.138*	.293*
Nepal	24	27.0	36.7	8.26	-.148*	-.094*
Poland	19	33.1	39.05	7.9	-.503	-.352*
Japan	16	32.5	41.64	9.25	.232	.248
Peru	14	32.2	33.6	7.6	.427	.619
BCSA sub-sample	394	29.8	B+	9.1	N/A	.354
Brazil	39	29.8	B+	8.12	.365	.335
Chile	37	29.1	B+	11.1	.522	.614
Ethiopia	22	29.6	A-	9.35	.379	.246*
Germany	15	28.3	A-	10.47	0.0 *	-.12 *
Jordan	15	31.7	A-	7.6	.025*	-.051*
Afghanistan	12	31	B	7.42	.277*	.167*
Peru	10	28.8	B+	8.9	.379*	.554*

* indicates not significant at 5% level

Of the 40 correlations reported 18 were non-significant - all of them involving groups with 26 or fewer members. However, of the 22 signficant correlations all except 5 were higher than the correlations obtained for the EPTB or BCSA sub-sample of which they formed a part. The larger correlations were obtained by the bigger groups, particularly in the EPTB sub-sample, with 6 above the 0.6 level. No definite cultural background could be detected although two of the countries were Arabic speaking (Egypt and Sudan). The third Arabic speaking country, Jordan, was represented by a much smaller group with a non-significant correlation. Two Spanish speaking countries showed high correlations (Mexico and Chile) and the third (Peru) also obtained high correlations, although one pair was non-significant.

The highest correlations were reported for the groups from Turkey, Thailand and Chile. Their average age was approximately that of the total sample and their EPTB/BCSA assessment means were close to the sub-sample means, except in the case of Turkey's which was higher. Their tutors' ratings means were not at the sub-sample's means. Those of Turkey and Chile were higher, while Thailand's was considerably lower. The Indonesian group was significantly older than the rest but still obtained satisfactory correlations, while the Algerian group was very young and with a high EPTB mean but low RT mean. This group was uncharacteristic in that they had been following a lengthy intensive English course in Britain when their English was assessed. This gave them some advantage in the listening subtests. Their promise was not carried over to their studies according to their tutors, and their courses were all first degree level.

The data for Chile and Brazil showed that both groups came with the mean BCSA ratings but varied by more than one SD in their tutors' ratings, with Chile's correlation significantly higher. Verbal reports from the British Council to the investigator had indicated that the assessments in Chile had been conducted more reliably than in Brazil. The data would seem to support this and also point to a difference in interpretation of the BCSA grading scale as represented by the identity of the BCSA mean grades and the discrepancy between the RT means.

Finally, Peru obtained high correlations for both measures, but the BCSA sub-group correlation was not significant.

6.4.2.2 The most significant factor to emerge from this data was that country of origin did not really seem to have an effect on the relationships between the two measures. But where a country group was large, ie numbering 30 or more, the relationship was positive and substantial as demonstrated by the correlations. This applied particularly to the EPTB sub-sample.

6.4.3 The sub-samples were next examined according to the subject area in which the students were working. It was thought that different subject areas might make different demands on the students' language proficiency and that this would be shown by variations in the tutors' ratings means. Where this proved to be the case, it was anticipated that correlations would be positive and significant and probably above the sub-sample mean.

6.4.3.1 Correlations for the groups were completed and are presented in Table 6.24 below, along with EPTB/ BCSA means and tutors' rating means.

Table 6.24 Correlations: EPTB and BCSA with R1B and RT by subject area

Subject area	N	DT	RT	r DT/ R1B	r DT/ RT
EPTB sub-sample	439	38.89	8.45	.319	.319
Agricultural, veterinary	41	37.5	8.6	.273	.155*
Arts	6	-	-	-	-
Medicine, nursing etc	47	38.9	9.1	.659	.503
Physical and biological sciences	73	38.8	8.8	.273	.298
Education, TEFL	39	39.9	8.4	.564	.498
Engineering, technology	81	38.2	8.3	.201	.278
English Studies, inc Applied Linguistics	20	44.5	7.2	.556	.616
Business, Social Sciences (professional)	63	37.9	7.9	.328	.267
Social Sciences (Academic)	68	39.2	8.6	.405	.504
BCSA sub-sample	394	B+	9.1	N/A	.354
Agricultural, veterinary	29	B+	9.1	.421	.334
Arts	21	A-	10.0	.131*	.088*
Medicine, nursing etc	44	B+	9.2	.114*	.261*
Physical and biological sciences	48	B+	9.7	.344	.252
Education, TEFL	59	A-	8.7	.63	.614
Engineering, technology	59	B+	8.8	.394	.408
English studies, incl Applied Linguistics	32	A-/A	9.1	.416	.435
Business, Social Sciences (professional)	61	B+	8.5	.236	.381
Social Sciences (academic)	95	A-	10.0	.173*	.172*

DT = EPTB Pt 1/BCSA total means * = non significant

Of the 34 correlations reported for subject area sub-samples all except 7 were significant beyond the 5% level, and all were positive. The range of significant correlations for EPTB was from .201 to .659, and the range for the ECSA sub-sample was .236 to .63. The group in the Arts subjects was too small to produce significant correlations.

6.4.3.2 Only the Education/TEFL groups obtained consistently high correlations (.5 to .6) across the two sub-samples, while students in the Physical and Biological Sciences sub-groups obtained correlations of around .3 in both sub-samples.

In the EPTB sub-sample the highest correlations were obtained in the following subject areas - English Studies, Education, Medicine and Social Sciences (academic). These subject areas could be said to be predominantly academic subject areas, along with physical and biological sciences, while the others contained a higher proportion of practical courses or components. These subject area groups were observed to contain the more proficient students in English according to the EPTB pre-departure scores. This feature was reflected in the BCSA total grades also but the BCSA sub-sample correlations in those subject areas differed from the EPTB sub-sample. The magnitude of the correlations of EPTB with tutors' ratings (above .4) in the academic subject areas suggested that EPTB was predicting more efficiently for the academic subject areas. Since the Battery was originally devised for students about to follow academic courses in Britain, these correlations provided evidence to support the validity of that claim.

Tutors' rating means were then examined with a view to establishing any differences between groups. The range of means was not very great - 7.2 to 9.1 for the EPTB sub-sample and 8.5 to 10.0 for the BCSA sample. The lowest mean (7.2) was obtained by the English Studies and Applied Linguistics group (EPTB sub-sample), while the equivalent group in the BCSA sub-sample had a rating at the sub-sample mean of 9.1. However, in the pre-departure assessments both

groups obtained the highest mean ratings - the EPTB group having a mean score one standard deviation above the sub-sample mean score. The inference drawn from this data was that a much higher standard of English proficiency was required of students studying these subjects. The evidence points to the fact that students in these groups arrived with a very high standard of English, which when applied to the needs of their particular studies was adequate, but little more. A further inference to be drawn was that students wishing to study English Literature, Linguistics or Applied Linguistics should be accepted only with very high grades or scores in their pre-departure assessments. These inferences are further supported by the fact that courses in these subjects require a high degree of sensitivity to the English language, even for native speakers. The correlations obtained support these conclusions also. For the EPTB group EPTB scores correlated at .556 and .616 with tutors' ratings, and for the BCSA group BCSA grades correlated at .416 and .435 with tutors' ratings. This strong relationship further indicated that tutors in a wide range of British institutions demanded a very high standard of English of students in these subjects before declaring a student "completely adequate" in his or her English proficiency.

It was therefore concluded that in the case of students of English Literature, English Language, Linguistics and Applied Linguistics the subject of study had a direct effect on the relationship between pre-departure and post arrival assessments of English proficiency.

6.4.3.3 The groups following courses in Education or TEFL were

found to be in a comparable position. This was in part to be expected since TEFL courses normally include a high proportion of elements found in Applied Linguistics courses. For both sub-samples pre-departure assessments were above the mean, at 39.9 and A-, and tutors' ratings were just below the mean, at 8.4 and 8.7. Moreover the correlations obtained were very high, .564 and .498 for the EPTB sub-sample and .63 and .614 for the BCSA sub-sample. These correlations seemed to support the notion that a high standard of English was required of students following courses in Education or TEFL.

6.4.3.4 Students in the subject groups Arts, Medicine, Physical Sciences, Social Sciences (academic) and Agriculture obtained tutors' ratings means above the means for both EPTB and BCSA sub-samples. These groups also obtained pre-departure assessment means at around the sub-sample mean, with the exception of the Agriculture group which had the lowest group mean on the EPTB. This could suggest that a lower standard of proficiency was tolerated for students in the agricultural and veterinary sciences.

Two other groups' performance were also examined, since it was commonly assumed that a lower standard of English would be sufficient for students in the subject areas of Engineering and Social Sciences (professional). In the EPTB sub-sample their EPTB means were 38.2 and 37.9 respectively. Their mean tutors' ratings were in all cases below the sub-sample mean, being the third lowest in the case of Engineering and second to lowest in the case of the Social Sciences (professional). In addition, there was a clear

difference in levels of proficiency attained by the Social Sciences professional and academic groups at both the pre-departure and the tutors' assessments. The academic group remained consistently higher at all times. One conclusion suggested by these observations was that a lower standard of initial English proficiency was not necessarily satisfactory for the Engineering and professional Social Sciences groups. If it had been, the tutors' ratings would have been higher and nearer the level given to the Social Sciences (academic) group. The EPTB Agricultural group provided an apposite contrast. Their group EPTB mean was 37.5, the lowest of the subject group means, but their tutors' ratings mean was third highest of the subject group means and equal to that of the Social Sciences (academic) group. This was taken to indicate that students in the Social Sciences (professional) and Engineering subject areas may require a higher standard of proficiency than is generally assumed.

6.4.4. The sub-samples were further examined in the light of the level at which they were studying. 87% of the sample were studying at the post-graduate level either for a Master's degree by tuition or for a research degree, or else were following a post-graduate or professional diploma (see Table 5.7 above). In each sub-sample students in the research degree group obtained the highest pre-departure scores and the highest tutors' ratings scores, while the Masters degree group obtained tutors' ratings at the mean and the diploma group ratings below the mean - see Table 6.25 below.

Table 6.25 Comparison of test and ratings means by level of study

	EPTB		BCSA	
	Pre-D	Tutors	Pre-D	Tutors
Sub-sample Mean	38.9	8.4	B+	9.1
Diploma level Mean	38.5	8.2	A-	8.6
Masters level Mean	38.8	8.3	B+	9.1
Research level Mean	40.3	9.4	A-	10.3

Although not too much could be deduced from these figures, it was noted that students pursuing research were considered to have adequate English proficiency above the mean, and students following diploma courses were more frequently rated as having some difficulty. There is an implication that language demands on students following diploma courses were likely to be greater than either they or their sponsors had anticipated. The strongest relationship between pre-departure and tutors' assessments was evident for the Masters level group with correlations of .43 and .48 (BCSA) and .32 and .3 (EPTB). For the EPTB sub-sample the relationship was even stronger for the Research degree group, with correlations of .42 and .4, a further indication of the Battery's power when predicting students in the more academic courses.

6.4.4.1 It was concluded that there was some effect on the relationship between pre-departure assessments and tutors' ratings, particularly with regard to the Research degree and Masters degree students, where the relationship was found to be stronger than with the sample as a whole.

6.5 Conclusion

The main conclusion at the end of the analysis was that subjecting the data to product moment correlations and to contingency tables was an appropriate method to be retained for the second investigation. In spite of the many reservations expressed about the nature and interpretation of such correlations, the majority of the correlations obtained pointed to a strong positive relationship between the pre-departure assessments and tutors' ratings. The strength of the relationship was confirmed by the contingency tables and supported by significant chi-squares.

On the basis of these results it was possible to give qualified confirmation of the hypotheses that significant and strong relationships existed between performance on EPTB and tutors' ratings of English ability and between grades obtained by British Council Subjective Assessments and tutors' ratings of English. These relationships provided evidence for the predictive validity of the two measures.

An analysis of the effect of other variables on the relationships had been carried out using principally the correlation method. It was observed that certain variables such as the following of pre-sessional remedial English courses and academic study in certain subject areas did affect the relationship. When subjects had followed pre-sessional courses in remedial English the relationship was seen to be less strong. The relationship was found to be strongest with groups studying in the more academic subject areas and Education and with groups either following Masters degree courses or undertaking research. There was no evidence that the cultural or linguistic background of the sample had a direct bearing on the relationship, but EPTB was most strongly

related to tutors' ratings in the case of large samples from Spanish speaking and Arabic speaking countries.

It was noted that subjects who had undergone EPTB could be said to have had a slightly lower proficiency in English on arrival in Britain than the BCSA sub-sample and that this difference was maintained, as revealed by tutors' ratings. The relationship between pre-departure assessments and tutors' ratings was of similar strength for both sub-samples although it was noted that the EPTB sub-sample produced stronger relationships for students pursuing the more academic courses and for large country groups. Although the numbers in each sub-sample arriving in Britain with inadequate English were reasonably low, there was evidence to show that the EPTB cut-off could have been lowered a little from 34.0 and that BCSA may have been excluding some borderline achievers who might have been able to study in Britain with success. One reason for this could have been that BCSA lacked the finer tuning afforded by the different sub-tests and total number of items in EPTB.

However, in order to ascertain that the findings reported were not particular to the sample of students whose English proficiency had been investigated, it was essential to conduct a similar investigation using a different sample and to establish the extent of support for the conclusions already reached. A second investigation was also needed to obtain a concurrent assessment of students' proficiency in English while in Britain through administration of the communicative proficiency measure.

CHAPTER 7 THE SECOND INVESTIGATION

7.0 The Purpose

The purpose of the second investigation was firstly to collect data on a new sample of overseas students in Britain through administration of the revised English Ability Rating form and the Communicative Proficiency Measure in order to confirm or reject the three hypotheses regarding the relationship between information obtained on the pre-departure measures and that obtained from the above "in study" English proficiency measures - reference paras 2.9.3, 2.9.4 and 2.9.5 above. Once the relationships had been established the second purpose of the investigation was to draw some specific conclusions on the predictive validity of EPTB and BCSA and some more general conclusions on the assessment of English proficiency for academic purposes.

7.1 The Method and the Instruments

The method adopted was identical to that of the first investigation, with the addition of the administration of the Communicative Proficiency Measure (CPM). There were four phases to the investigation:

(i) identification of the sample and collection of background data, including details of EPTB scores or ECSA grades obtained,

(ii) administration of the English Ability Rating form to subjects' tutors,

(iii) administration of the Communicative Proficiency Measure to a sub-sample,

(iv) collection and analysis of results.

7.1.1 The timetabling of the phases was as follows:

November - collection of sample background data

December - collation and coding of data

January, February - Distribution of English Ability rating forms to tutors for completion and return

March, April - administration of Communicative Proficiency measure

May, June - collation and coding of results.

7.1.2 The background data was collected from the personal files of students constituting the sample in the British Council Headquarters in London, as in the first investigation. A 'Personal Data and Test Results' sheet was used. Data for eight subjects was recorded on each sheet and all items of information were coded for transfer to punch cards ready for computer analysis. A copy of the data sheet is reproduced at figure 7.1, together with a key to the symbols denoting each item of information.

Figure 7.1 Personal Data and Test Results Sheet

Name	....	Name of subject
Sex	....	Sex
Age	....	Age
Coun	....	Country
Qual	....	Highest qualification
T1/L	....	Test 1 score/Listening grade
T2/S	....	Test 2 score/Speaking grade
T3/R	....	Test 3 score/Reading grade
T4/W	....	Test 4 score/Writing grade
Tot	....	EPTB Pt 1 Total/BCSA 'total'
5	....	Test 5 Score
O	....	Oral grade/as T2/S
W	....	Oral grade/as T4/W
Rem E	....	Weeks of remedial English
Univ	....	British institution attended
Dept/Subj	....	Subject area
Level	....	Level at which studying
Length	....	Length of course

T1 to T5 denoted the EPTB sub-test scores, while /L to /W denoted the BCSA skills gradings. T5, the Reading Speed test in Part 2 of EPTB was not completed by all the EPTB sub-sample and was not applicable to the BCSA sub-sample. Grades at O and W denoted grades awarded to students who had taken EPTB and who had been given supplementary oral and writing tasks. Grades obtained by BCSA sub-sample subjects at /S and /W were repeated alongside O and W for convenience of analysis.

Only three changes were made from the data sheet used in the first investigation (see para 5.1.1 above). The mother tongue of each subject was not recorded for the second investigation. Sub-samples according to mother tongue had not been analysed in the first investigation. The large country samples eg Thailand, Mexico, Sudan were almost all monolingual, and where a large country group was multi-lingual, eg Indonesia, many subjects had recorded the official language and not necessarily their own mother tongue. Most other multi-lingual groups tended to come from African or European countries with small numbers. There were thus up to 18 items of information recorded for each subject.

For the second investigation the pre-departure and pre-course data were recorded first, and the last four items recorded related to subjects' studies in Britain. The third change related to the T1/L to T5 codes which were more explicit than the single figures used in the first investigation. Data was collected on 972 subjects.

7.1.3 English Ability Rating forms were prepared for each of the subjects. However, by the time the forms were ready for

distribution, some subjects had left their courses or could not be traced, and others turned out to be applicants whose applications had been accepted but who did not in the end take up their places. The final total of rating forms sent out was 925.

7.1.3.1 The major change that had been made to the English Ability Rating form was the inclusion of a fourth option in section A of question 1 (see discussion at the end of para 5.4.6.1 above) as a result of tutors' comments. The four options in response to the question "To what extent is the student's general ability in English now adequate..." were

- completely adequate
- adequate
- only just adequate
- inadequate

In general terms it was felt that the two middle categories would fill the large gap between the extremes of <u>completely adequate</u> and <u>inadequate</u>, especially as many tutors had clearly felt that some students were far from adequate but not really so lacking in proficiency as to be deemed inadequate.

The siting of section A on the form was modified. Instead of setting out the options vertically, alongside the options for section B, the four options were placed across the page above section B. In this way it was not possible for respondents to relate categories in section A with categories in section B on the basis of their relative positions on the page.

Two minor additions were effected in the definitions in the options in question 1B. References to pronunciation had not been included

previously, but some tutors had pointed out that pronunciation, deviant or otherwise, was often a key feature of a highly proficient speaker's English. Similarly references to the subject's control of English usage had been made only in the definitions of the three lowest categories of proficiency. These omissions were made good in the first two definitions (added elements underlined).

- Shows native speaker ability in English usage

- Clearly a non-native speaker because of minor faults in English usage and pronunciation, but this does not handicap him/her in his/her studies.

In the accompanying key tutors were advised to delete the words and pronunciation if they considered they were not applicable.

As a result of the change to section A of question 1, the labelling of the skills rating scales in question 2 was amended to include the four categories. The length of the scale remained the same - 12 cm.

7.1.3.2 In questions 3 and 4 October 1973 had to be deleted and the words beginning the course of study were substituted. October 1974 was not included because by no means all the sample had begun their courses in that particular month.

A fourth option, not applicable, was added to question 3 (relating to improvement in English ability) with a note in the key to the effect that the question might not apply to students with a very high standard of English on arrival. It was hoped that students in this situation would be separated from those who were deemed to have made no improvement, although they had not arrived with a high standard of English.

Question 4, relating to whether students had received tuition in English since the beginning of the course, was simplified to two options - Yes and No, don't know or not applicable. The only useful information here was the number for whom the response was yes. Splitting the response no into several categories served no purpose.

Question 5 was not wholly open-ended, as previously. The qualification aimed at was requested, but two further options of on attachment only and research only were given to cover the cases where no specific qualification was being aimed at. Raters were again asked to use the back of the form for any further comments they wished to make (question 6).

7.1.3.3 A new letter to the tutors who were requested to complete the rating forms and a new letter to the British Council regional directors who were requested to distribute the forms to the appropriate tutors were drafted. Copies of these letters and of the English Ability Rating form and the key used in the second investigation are to be found in Appendix III.

7.1.4 The development and trialling of the Communicative Proficiency Measure were presented and discussed in paras 4.6.1 to 4.6.9 above. At the end of the trialling it was decided that the measure should consist of three components:

(i) Reading test comprising 4 cloze passages

(ii) Essay using one topic for all

(iii) Interview following a fixed format.

Copies of the different sub-tests used in CPM are given in Appendix III, but brief descriptions of the different components follow below.

7.1.4.1 The reading test consisted of four cloze passages as follows:

Passage	Items	Topic	Difficulty
AC	30	General science	37%
CC	30	Mathematics and education	38%
DC	25	General interest (history)	37%
MC	24	Sex control technology	50%
Total Items	109		

Each passage was preceded by an introductory paragraph or sentences before the deletions started. Passages AC, CC and DC were systematic cloze passages with every sixth word deleted while passage MC was a modified cloze passage with only structural words deleted, but at the rate of one deletion per five words of text. The degrees of difficulty indicated above for each passage were very approximate and based on the results of the very small sample who trialled the passages (see para 4.6.8.4 above).

Subjects were instructed to read as much of each passage as possible in 30 minutes, to write one word in each blank, to fill as many blanks as possible, and to attempt the passages in any order. A short example consisting of one sentence with three blanks was provided. The full instructions for the candidates and the texts are reproduced in Appendix III.

By limiting the total time spent to 30 minutes, students were

encouraged to work fast. Since they were likely to find some passages more interesting and probably easier than others, candidates were encouraged to work them in the order that they wished to. In this way it was hoped that students would be able to give their best performance in the limited time available and not waste time wrestling with the more difficult passages. By limiting the time, the more able candidates were thought likely to obtain much higher scores relative to the less proficient, and often slower, students. 1 point was awarded for each correctly filled blank and only the original words in the source texts were accepted as being correct. The total maximum was 109 marks.

7.1.4.2 The writing test was administered immediately after the reading test. The purpose of the essay writing was to elicit a sample of writing from each subject in which were displayed ability to write narrative or descriptive English and to write a paragraph or two of discursive English. The instructions were geared to achieving that purpose.

The single topic was the same for all candidates but sufficiently autobiographical and general so that every student would be able to find something to write about. The topic was essentially the same as that used in the trialling - describing and discussing part of one's work that had been found to be interesting. This was expanded to include an aspect of living in Britain as an alternative to writing about work, in case the candidate should feel unhappy about writing on his work - particularly if he or she had found it disappointing! In the trialling of the essay three paragraphs

of suggestions and guidance had been offered to the students, but they had largely been ignored. These suggestions were therefore dropped from the final CPM. Instead, a short paragraph in brackets preceded the topic, in which the investigator attempted to set the context for the writing and to indicate that there was some purpose for the writing.

Some general guidance in the form of notes was, however, given so that students should be aware of certain parameters within which they were to write. They were advised to be brief in view of the time limit; they were advised that factual accuracy was not as important as fluency and style, and were told that the marker was looking for information and opinions. In this way it was hoped that it would be clear to each candidate what the purpose of the writing was and the principal elements to be written about in the limited time available. The time limit was clearly stated.

The criteria adopted for assessing the essays were overall criteria on a six point scale, as decided at the end of the trialling (see para 4.6.8.6 above). The original definitions used in the trialling, see para 4.6.5.3 above, were amplified for the second investigation in order to make them more explicit and to bring them closer to the revised definitions of the oral interview criteria and to the definitions in Question 1 section B of the rating form. The main changes were to introduce parameters of communication - whether it was established and the degree to which it had been established, accuracy of usage and appropriateness of style. The criteria were:

Level 6. Near native speaker ability (with or without a few minor faults).

5. Some minor faults (probably in style, choice of lexis or occasional points of grammar), but fluent and generally very competent in communication of information and ideas.

4. Competent in communication of information and ideas, but displaying some weaknesses in style or use of vocabulary or grammatical accuracy. Not always very fluent but adequate.

3. Serious weaknesses are evident, and at times the successful communication of information or ideas is threatened. Barely adequate for academic writing.

2. Serious deficiencies in the control of grammar, sentence structure and lexis limit the effective communication of information or ideas. Inadequate.

1. Little or no ability to communicate effectively in written English.

In view of the similarity of the information obtained by both the overall and analytic marking in the trialling of the essay (see para 4.6.8.3), it was decided not to give separate marks for different features of the writing but to keep a checklist of

organisation and relevance of the content,
appropriateness of the expression, and
accuracy of the expression

available which was consulted each time an overall level was awarded.

7.1.4.3 The procedure used for the interview during the trialling had proved satisfactory and no basic change was made for the investigation. However, one extra phase - or topic of discussion - was added, and the fourth phase was split into two phases for ease of description and administration. The phase that was added (phase 4 below) concerned the subject's own ideas of his abilities in English. It was felt that since the whole enquiry was concerned

with the assessment, improvement and adequacy of students' English while following courses in Britain, subjects should be given an opportunity to air their views on the matter. Candidates were therefore questioned on how their English was assessed in their home country, on any English courses they had followed at home or since arrival in Britain, and on their own assessment of how satisfactory their English was for their purposes. In many cases interesting discussions were held on the quality and suitability - or lack of them - of English language courses that had been followed on arrival in Britain. The two final phases of the interview gave the students the opportunity to ask questions, often a natural consequence of the discussion on English proficiency, and drew the interview to a close.

The different phases and conduct of the interview - with the exception of the new phase discussed above - were described fully in paras 4.6.3 and 4.6.8.5 above. An outline of the phases and their content is given below, with fuller details on questions and topics in Appendix III.

Phase 1. Introductory

- statement by interviewer
- setting at ease with checking of some administrative and personal details
- transition to next phase by asking about family or mail service or personal welfare.

Phase 2. Personal background

- home country
- travel to Britain
- accommodation - discussion
- life in Britain - discussion

Phase 3. Work being undertaken

- details of course
- description of typical day
- description of project/research - discussion
- discussion of value on return to home country

Phase 4. English language ability

- assessment in home country
- courses followed at home or in Britain - discussion
- problems encountered
- self-assessment and improvement - discussion

Phase 5. Questions

- any questions that subject may wish to put and answers/discussion from interviewer

Phase 6. Closing the interview

- choice of moment to close
- expression of thanks by the interviewer.

Each of these phases was to be followed although it was clear that in many cases discussion at one stage might be prolonged and another stage or two might have to be treated quickly. However, the procedure decided upon gave ample opportunity for the subjects to give the interviewer a lot of information of which he would be quite ignorant, and in each of the major phases there were opportunities to develop discussion and argument, thereby extending the range of language the candidate would have to use. A limit of 20 minutes was set by the interviewer, although it was found that this time was often varied.

During the trialling of the interview two approaches for assessment

of the oral performance were used - analytic and impressionistic. As a result of the trialling it was decided to use only the overall impressionistic criteria, on a 6 point scale, with improved definitions for each point on the scale. The criteria used for the trialling are given in para 4.6.3.3 above. The amended definitions given below have added the parameters of communication, fluency, grammatical accuracy (usage) and pronunciation.

Level 6. Native speaker ability and fluency with (occasional) evidence of non-native speaker pronunciation features.

5. Full communication established, but with occasional minor faults of English usage and pronunciation.

4. Adequate communication, occasionally impaired by a number of minor faults of usage or pronunciation. There may be occasional hesitations.

3. Communication established but impaired from time to time as a result of frequent faults in grammar/ lexical choice/pronunciation. There may be hesitations and expression may come in short or disjointed phrases.

2. Frequent limitations in communication by serious deficiencies in grammatical control and pronunciation.

1. Hardly any or no ability to communicate.

It was anticipated that some subjects might exhibit characteristics of performance at more than one level eg 3 and 4. When this was the case, it was decided that the dominant level should be attributed.

7.1.4.4 The measures had been devised in such a way as to retain maximum flexibility of administration. In the case of large groups more than one person was needed for the administration, but it was anticipated that in most cases the groups would not be large, subjects would not all arrive at one time, and the investigator

would be the only person to supervise the written and reading tests as well as conduct the oral. There was no obligatory sequence, although it was felt that the oral would be most satisfactory if conducted at the end, as the subjects would have settled in thoroughly by that time and would probably need a much shorter period of 'warm up'. It was also thought that the reading sub-test would best be done first. In the cloze passages the student had to interact with given text in a formal way whereas the essay and the interview required the ability not only to produce but to think of and organise the material to be communicated.

Accordingly three types of situation were anticipated. The first was the easiest and only possible with the cooperation of host departments or institutions. It entailed administration of the reading and essay tasks to a body of subjects at the same time, to be followed immediately - or at set times later - by individual interviews conducted by the investigator. This method involved the subjects in waiting for their interviews and required either a considerable block of time when no lessons were timetabled or the use of the subjects' free time.

The second situation envisaged was the 'walk in' administration when subjects were given a period during which they were asked to come and take the tests. The exact time of arrival depended on their personal timetable on the day in question. Students could arrive individually and begin the reading and writing tests immediately. On completion they then had the interview. In this way not much time was lost and individual students did not need to wait too long for the interview, if at all. The speed of the

procedure depended on the volume of students attending. The major drawback was that the interviewer had to interrupt an interview briefly every time somebody arrived to do the test in order to hand out the papers and check that instructions were understood.

The third situation anticipated was a combination of the other two and occurred if for some reason most of the subjects to be tested arrived at the same time! In that case all but one began the written tests together, and the remaining one was interviewed first. At the end of the interview a second subject was interrupted in his work and taken aside for interview, after which he resumed the written tests. This method of interrupting the subjects was not desirable but had to be considered if the situation warranted it. In the event all three methods of administration were employed, and no major difficulties arose, except that some students did not have the time to wait for interviews.

The arrangement of the testing sessions is discussed later in the chapter, but it was anticipated from the outset that the investigator would either have to contact directors of courses to arrange for testing sessions for specific groups of students or else write to individual students requesting them to attend at certain times in a specific place. In the former case direct correspondence and telephoning were employed to make the arrangements. In the latter case a form letter was prepared explaining the purpose of the enquiry and the need for the testing, requesting the subjects to give of their time and to cooperate. A copy of the form letter sent is given with the copies of the CPM in Appendix III.

7.2 The Sample

The sample was chosen on the same basis as the sample for the first investigation, ie students who

(a) had begun their main study in Britain during the previous 3 months, and who

(b) would be spending a minimum of 6 months on their course of study, and who

(c) had had their English proficiency assessed either by EPTB or BCSA before their departure from their home countries.

The sample was selected in early November and so most subjects had begun their principal course of study or training during September or October. A number had been in the country for considerably longer, however, because of undergoing extensive periods of pre-sessional English language training.

925 subjects were identified for the sample, and background data was obtained on each subject. English ability rating forms were sent out to each subject's tutor, and a sub-sample of 95 took the Communicative Proficiency Measure. However, the rate of response by the tutors was not so high for the second investigation, possibly because the novelty of the first assessment no longer obtained or possibly because the request came earlier in the second term with ample time for completion. Some tutors may have set the form aside for a couple of weeks and then forgotten to complete it. One tutor in charge of a group of francophone African teachers gave the forms to the students to complete, and requested that no further forms should be sent! Subjects for whom rating forms were not returned by tutors and subjects who had assessed themselves were not included in the finalised sample, which totalled

729. This total represented a 79% response from tutors and a 12½% reduction in size of sample compared with the first investigation.

7.2.1 Characteristics The sample for whom data was analysed consisted of 729 subjects from 82 countries. Only 117 (16% of the total) were female. Characteristics of the sample's background are reported briefly in the paragraphs below. Detailed tables on the characteristics and the assessments of the sample are given in Appendix IV. References to these tables are put in brackets in the text.

7.2.2 Age The ages of the subjects ranged from 18 to 50, with a mean age of 30.3 yrs - almost identical with that of the first investigation sample, which was 29.9 yrs. (Table 7.2 in Appendix IV.)

7.2.3 Countries of origin Compared with the first investigation numbers increased from Africa (South of the Sahara) and from the Middle East and North Africa. There was about a 10% decline in numbers from the other areas. One of the features of the sample was the increase in subjects from Commonwealth countries, particularly in Africa, who accounted for 9% of the sample from 14 countries. Large groups were tested in Zambia and Botswana. The region providing the largest group of students (28%) was again Latin and Central America. The largest single country groups again came from Mexico, and Sudan, Brazil and Thailand also provided large groups.

179 subjects - 24.5% of the total sample - came from 17 Spanish speaking countries, while 158 - 21.7% of the total sample - came

from 12 Arabic speaking countries. Fuller details are given in Tables 7.3 to 7.5 in Appendix IV.

7.2.4 Educational background Whereas approximately three quarters of the sample had first degrees or second degrees, approximately one fifth had not proceeded beyond the end of secondary edcuation or its equivalent. That represented a substantial change from the first sample, of which only one twelfth had not continued beyond the secondary stage. There was a reduction from 22.4% in the first investigation to 15.9 in the second of those who had already obtained second degrees. (Table 7.6, Appendix IV.)

7.2.5 Studies in Britain The sample was distributed over 134 institutions or organisations for their studies in Britain. Almost all were academic institutions, although a few students were undergoing attachments with private firms. The largest group of students was studying at colleges and polytechnics in the greater London area but with concentrations at only a small number. In the rest of the country the main concentrations were in Manchester, Edinburgh, Oxford and Leeds.

7.2.5.1 Study areas were categorised into nine main main areas with a tenth, miscellaneous, category. Between a fifth and a quarter of the sample were engaged in Engineering or Technological studies. This represented a major increase - from 16.8 to 21.8% - when compared with the first investigation sample. Similar increases were observed in those studying Education subjects - from 11.8 to 16.2% - and Social Sciences (professional) subjects - from 14.9 to 18%.

There were corresponding decreases in the proportions studying Social Sciences (academic) subjects, Medicine and the Pure Sciences.

7.2.5.2 Approximately 80% of the sample were studying for Master degrees, professional or academic diplomas, or pursuing research. The major change over the first investigation sample was that 6.8% fewer subjects were doing research or attempting research degrees. There was an increase of 4.5% in academic attachments and a slight increase in the numbers studying for first degrees. Fuller details, with comparisons with the first investigation sample, are given in Tables 7.7 to 7.9 in Appendix IV.

7.3 Pre-departure English Assessments

316 subjects (43.5% of the sample) took EPTB before departure from their home countries. 413 (56.5% of the sample) had their English assessed according to the British Council Subjective Assessment procedures.

7.3.1 The overall mean obtained by the EPTB sample on Part I was 38.32, 0.6 lower than the previous sample and 1.68 points lower than the standard mean. This low mean for the second year running provided more evidence to support the impression that the general standard of English proficiency of officially sponsored students from overseas was declining slowly. 130 took the Part II speed reading test, obtaining a mean score of 66.2. This score was well below the previous sample's mean score of 76.0 and below the standard mean of 70. However, some of the sample may have taken the new version C of EPTB which had fewer items in Part II, and for which the mean was 62.

One sixth of the subjects obtained Part I scores of less than 34.0, the cut off score below which applicants are normally advised to follow several months of part-time English classes and then try the test later! However, these particular subjects managed to reach Britain to study and were valuable for the current investigation since they represented low achievers who might not normally have been expected to be available. The size of the group with inadequate English was proportionately larger by 50% than the equivalent group in the first investigation.

Approximately half of those taking EPTB took oral and writing tests to supplement their EPTB scores. A mean of a low B+ was obtained for the oral and a low B for the writing test.

Performance in the oral was generally better than performance on the writing tests, as in the previous sample, with the mode at A for the oral and at B for writing. Grades were reported with pluses and minuses, but it can be assumed that there was little or no difference between a plus and minus grade eg between C+ and B-. (Tables 7.10 to 7.12, Appendix IV.)

7.3.2 Most subjects who were assessed using the BCSA were graded separately for listening, speaking, reading and writing. However, in a few cases grades for the individual skills had not been given and only an overall grade, or sometimes a statement, was given. Although no total grade was normally given to applicants, notional 'total' grades were computed for each candidate for the purposes of the current investigation.

Although grades were not reported with minuses for the individual skills, the mean grade for listening was reported as A- as it was marginally better than B+. The mean grades were identical with those obtained for the first investigation sample, but the distribution of 'total' grades varied a little, with the second investigation sample being slightly less proficient. Whereas total grades A and B+ accounted for 68% of the first sample they only accounted for 54% of the second sample. It was thus concluded that both the second EPTB and BCSA sub-samples were slightly weaker in their pre-departure English assessments than the first sample equivalents. (Tables 7.13 to 7.14, Appendix IV.)

7.3.3 The only measures in common for both EPTB and BCSA sub-samples were the oral and writing test grades. In the first investigation the BCSA sub-sample had obtained means nearly a half grade higher than the EPTB sub-sample. However, in the second investigation the two sub-samples were much closer in mean ability. The EPTB sub-sample obtained B (low) for Writing abd B+ (low) for Oral and the BCSA sub-sample B and B+ (low). (Table 7.15, Appendix IV.)

7.3.4 Just over 60% of the sample followed pre-sessional courses of remedial English. This was nearly 5% higher than in the previous sample. The main increase was in members of the BCSA sub-sample. The mean length of tuition was 8.5 weeks, with a range of 2 to 26 weeks. Length of courses was distributed evenly but the BCSA sub-sample spent slightly more time on such courses with a mean of 8.9 weeks, while the mean length of tuition for the EPTB sub-sample

was 7.9 weeks. (Tables 7.16 and 7.17, Appendix IV.)

7.3.5 As in the first investigation, language performance and improvement data were analysed in three different ways - for the whole sample where possible, for the EPTB sub-sample, and for the BCSA sub-sample. Comparative characteristics for the whole sample and for the two sub-samples are given in Tables 7.18 to 7.23 in Appendix IV.

The age distribution and mean of both sub-samples were almost identical, and the EPTB sub-sample contained proportionately fewer female subjects. Educational background also followed the same distribution pattern.

The chief difference between the sub-samples lay in the countries of origin. 59% of the EPTB sub-sample came from Middle Eastern and Asian countries, while only 49% of the BCSA sub-sample came from these two regions. Since a further 24% of the EPTB sub-sample came from S American countries, it contained relatively few representatives of N European and African countries. As a consequence the the BCSA sub-sample contained a higher proportion of subjects from African and European countries. Nevertheless it was concluded that except with respect to certain countries of origin there was no significant difference in the background characteristics of the two sub-samples. As with the first sample the largest country groups, with the exception of Brazil, had been assessed by EPTB.

7.4 English Ability Ratings

The response of tutors was again very satisfactory although the resultant

sample was smaller for the second investigation. Tutors apparently had no problems with completing the questionnaire. The main problem was finding time when one particular tutor had a large group of students. If there were serious problems with completing the questionnaire, these were not made known to the investigator.

It was not possible to ascertain why some tutors had not completed and returned questionaires. The cooperation of the tutors and of the regional British Council offices staff was again much appreciated.

7.4.1 The first question answered by the tutors was,

> "To what extent is the student's general ability in English now adequate for undertaking his/her studies, research or training (ie as at Jan/Feb 1975)?"

Answers to that question are summarised in Table 7.24 below.

Table 7.24 Tutors Rating 1A

		Frequency	Frequency %	Cumulative Frequency %
1.	Inadequate English	16	2.2	2.2.
2.	Only just adequate	157	21.6	23.8
3.	Adequate	372	51.2	75.1
4.	Completely adequate	181	24.9	100.0
		726	100.0	

Only 2.2% of the whole sample were judged to have inadequate English at that stage in their stay, but just over 21% of the sample were judged to possess English ability that was only just adequate. The inclusion of this category in the rating form seems to have been justified as many of the tutors clearly had strong reservations about their students' English. Since the categories in the response to question 1A had been changed it was not possible to compare the

performance of this sample with the first investigation sample, except for the category of inadequate. Responses at 2.2% for the second sample compared with 4.1% of the cases in the first sample.

Tutors were given a second set of responses to the same question. This time six definitions, summarised in Table 7.26 below, were offered. Only one student was deemed totally inadequate in his English but another 43 were deemed to have considerable deficiencies and to require more proficiency. This second category clearly spanned the borderline area between inadequate and only just adequate. A Pearson product-moment correlation between the two sets of responses was computed at .776, showing a strong relationship. Distribution of responses 1 and 2 in 1A seemed to correspond closely with responses at 1, 2 and 3 in 1B. Responses to question 1B paralleled the responses to the same question in the first investigation, as shown in Table 7.26 below.

Table 7.26 Tutors Rating 1B

		Frequency	Frequency %	1st investn Frequency %
1.	Totally inadequate	1	0.1	0.4
2.	More profy essential	43	5.9	4.1
3.	More profy desirable	132	18.1	18.6
4.	Minor deficiencies	159	21.8	23.3
5.	Adequate	351	48.2	46.7
6.	Near native speaker	42	5.8	6.9
		728	100.0	100.0

7.4.2 In question 2 tutors were asked to assess the subject's ability in each of the four language skills by putting an X at the

appropriate point along the scale shown in Figure 7.28 below. (Tutors did not, however, have numbers or divisions on the scales they received. They were added later for scoring the responses.)

Figure 7.28 Question 2, skills rating scale

	completely adequate		adequate		only just adequate		inadequate
	13	11	9	7	5	3	1
Ability to understand spoken English							
Ability to speak English							
Ability to understand written English							
Ability to write English							

The length of the scales was 12 cm. For ease of calculation the extremities of the scale were attributed 1 and 13. Only if crosses were put at the very extremities were these points awarded. In two cases 13 was awarded, and the highest mark was otherwise 12.5. 2 was the lowest point on the scale used by the tutors, except for one case of 1.

66.5% of the tutors differentiated between skills and did not mark ability as being equal in all skills. As a group, tutors deemed their students most proficient in the reading skill, closely followed by listening. Speaking and writing were deemed to be the skills in which students were least adequate. The difference between the means for reading and writing was 1.4, or more than 10% of the scale. The modes for all the skills was consistently in the adequate band of ratings, although for reading the

completely adequate group was almost equal in size. Tutors discriminated most in the writing skill with over 4% of the sample assessed as being inadequate and 23% as only just adequate. In contrast, less than 9% were adjudged to be only just adequate or inadequate in reading. (Table 7.29 and Figure 7.30 in Appendix IV.)

Further evidence of the discrimination made between ability in the different skills by tutors was obtained by intercorrelations. Pearson correlation coefficients obtained were:

Listening with Speaking .72, with Reading .68, with Writing .59
Reading with Speaking .63, with Writing .66
Writing with Speaking .67

(Table 7.31, Appendix IV)

7.4.2.1 Tutors were not requested to give an overall skills rating in question 2, since they had made general responses in question 1. However, the investigator summed the four individual skills ratings to produce a notional 'total' skills rating on a scale from 4 to 52. This scale was considered another form of assessment of general ability - the sum of the skills adding up to the whole - and compared with the ratings at question 1 and also with pre-departure assessments. Total ratings were not computed unless all four skills had been rated. They ranged from 16 (borderline between inadequate and only just adequate) to 50 (completely adequate). For the purpose of categorising the scores, the precise layout of the scales was taken into consideration. Thus the mid point between the lower two categories was taken as one cut-off, the first point clear of the words only just adequate provided the next - it was also the mid point on the scale - and the final point was to the right of

the y of completely adequate, as in figure 7.32 below. The points on the scale where the cut-offs occurred were at 4.2 (Total = 17), 7.2 (29), and 10.8 (43).

Figure 7.32 Skills rating scales - cut-offs and

completely adequate	adequate	only just adequate	inadequate	
	10.8	7.2	4.2	individual
	43	29	17	total

The mean total rating for the sample was 37.3 (centre of adequate division), with standard deviation of 7.4 (15% of the scale).

7.4.2.2 Tutors' responses to question 1 of the rating form were analysed according to sub-sample. Members of the EPTB sub-sample were consistently adjudged to be a little less proficient than the BCSA sub-sample although according to responses in section B the range of proficiency of the BCSA sub-sample was greater. 30% of the EPTB sub-sample were assessed as only just adequate or inadequate compared with 20% of the BCSA sub-sample, while 31% of the BCSA sub-sample were assessed as completely adequate, compared with only 17% of the EPTB group. (Tables 7.25 and 7.27 in in Appendix IV.)

The difference between the sub-samples was also expressed in the skills ratings in question 2, with the skills rating means of the EPTB group between one third and one half SD lower (except for writing), and with BCSA distributions skewed more negatively. (Table 7.33 in Appendix IV.)

Differences followed a similar pattern to those of the first investigation sample (see table 5.23) although the ratings means were lower and standard deviations greater for that sample.

7.4.3 The third question on the rating form related to improvement noted in the students' proficiency since the beginning of the main course of study. A new response - not applicable, because of a very high initial standard of proficiency - was included for the second investigation. Tutors' responses showed that a fifth of the total sample were adjudged to have made <u>considerable</u> improvement, while just over half had made a <u>little</u> improvement. These responses differed significantly from those of the previous year when nearly a third were adjudged to have made <u>considerable</u> improvement and almost two thirds a <u>little</u> improvement. (Table 7.34, Appendix IV.)

It was not clear why the difference between the two sets of responses was so great. Three reasons were suggested. The students may well have begun their courses with a higher standard of English than the members of the first sample. This inference found support in the fact that a higher proportion of the second investigation sample had followed remedial English classes and for an average of a week longer. It could also have been that as the tutors were responding during the first half of the second term and not towards the end of that term there may not have been time to have noted all those who had made <u>considerable</u> improvement. Finally there had been 20% missing responses in the first investigation, many of which were probably missing because of the lack of the fourth category of response.

Responses for the two sub-samples were almost identical, the major difference being that it was considered that improvement was not applicable (category 4) to 14.5% of the BCSA sub-sample but to only 9.9% of the EPTB sub-sample. (Table 7.35, Appendix IV.)

7.4.3.1 About a quarter of the total sample were known by their tutors to have had some language tuition since starting their study. This compared with 21.7% in the first investigation. However, the accuracy of these responses can be questioned since a number of tutors indicated that they were aware that particular students had had pre-sessional English. Question 4 was intended to relate solely to English tuition since the beginning of the main course of study. (Table 7.36, Appendix IV.)

7.4.3.2 Question five was again included to provide the latest information as to the level of the course of study being followed by the student. This information was incorporated into the data reported in paragraph 7.2.5 above.

7.4.4 Tutors were invited to make further comments on the reverse of the rating form. Comments were made on 165 (22.6%) of the students. Some consisted of only one sentence whereas others consisted of a number of paragraphs. 30 did not concern language proficiency. The remainder did and are reproduced in Appendix IV, section 5.

Comments were analysed and classified into five general topic areas:

- Further explanation of responses given

- Further comments on the individual student
- General comments on the English of students from overseas
- Comments on the organisation of the questionnaire
- Other comments, including the British Council.

7.4.4.1 About one third of the comments amplified the responses given in particular sections of the rating form, dealing with specific language abilities, with improvement in proficiency, or with the following of English courses. The following example covers both types of comment and was received in respect of a male student, No 400, from Turkey.

> Q1B. Spoken English - He makes occasional errors with words with non-native pronunciations.
>
> Written English - His grammar and vocabulary are good but he occasionally uses non-technical words in place of more appropriate medical terms.
>
> Q4. He has taken a course in English with the British Council, but it was not sufficiently advanced. He plans to take, or may currently be undertaking, a more advanced class.

Other examples of this kind of comment can be found for subjects 393, 539, 840 and 998.

The major areas of difficulty for students were identified by tutors as:

reading - difficulty coping with the volume of required reading, and the consequent need to develop speed;

writing - the need to write effectively within strict time limits, as in an examination;

listening - the listening load for some, because of the volume of lectures, and difficulty in getting accustomed to regional accents;

speaking - the need for discussion skills.

7.4.4.2 The greatest number of comments (over 40%) were classified as 'further comments on the individual student'. They differed from the above in that they did not normally relate to specific sections of the questionnaire but were considered by the tutors to be information relevant either to the performance of the subject being assessed or to the investigation as a whole. Comments covered a student's ability in his or her speciality, aspects of personality, such as shyness, reticence, positive attitude, persistence, and, in two cases, the fact that students had fellow Arabic speakers in their class with whom they could communicate when they had had difficulty understanding lecturers. In a number of cases British staff had worked with students in their home countries and so felt that this extra experience and knowledge of the student's background helped overcome any communication difficulties that may have arisen. Other students had already worked with English-speaking technical assistance personnel on projects before they left for Britain.

7.4.4.3 Nearly 20% of the remarks analysed commented on more general topics affecting the English proficiency of overseas students. Some tutors commented on the fact that difficulties lay more with social English and not the use of technical English required for the course (see comments on subjects 525, 643). Another comment was that it was not so much language which was the problem but understanding the English approach to many disciplines eg law (see subject 727). Others noted that the student had high intellectual ability which in their opinion would compensate for and overcome any linguistic problems - one tutor anticipating that 'because of his high intellectual ability I suspect that his

comprehension of English may soon overtake native born Britishers' (see comment on 645). A number of tutors made comparisons with British native speakers suggesting not only that in some cases non-native speakers would become more proficient but that in many cases native speakers experienced similar or identical problems, eg in comprehension. Cultural adaptation, or 'cultural strangeness' as one tutor put it, was also highlighted (eg subject 1784), and at least two tutors mentioned the fact that students who had brought their wives to Britain gave themselves much less opportunity to speak English with other students (eg subject 914).

Comments listed under other comments were very few and often related to a student's background history of preparatory English courses (subject 538), the nature of British Council language assessment overseas (subjects 815 and 522), the importance of serious language assessment (subject 1016).

7.4.4.4 Seven comments were made on the questionnaire (subjects 75, 281, 335, 355, 1399, 591, 2100). One was by implication derogatory, another complimentary. Others found difficulty with general classifications beyond <u>adequate</u> (question 1A). One tutor felt the gap between <u>adequate</u> and <u>completely adequate</u> was too great and proposed a further category <u>more than adequate</u>, while another tutor considered <u>outstanding</u> to be preferable to <u>completely adequate</u>. Two tutors suggested that 'usage' and 'pronunciation' should be separated (question 1B). They may not have read the key and so not understood that 'pronunciation' could have been deleted by them. However it may have been that they would have wanted to

delete 'usage'. This could have been made explicit in the questionnaire.

7.4.4.5 Although there were fewer comments by tutors than in the first investigation, they were very relevant and in many cases of positive help to the investigator. The questionnaire seemed to be very satisfactory in its present form and no major complaint or adverse criticism was expressed. The most frequently expressed reservation was that of time - either of having insufficient time to complete a group of forms or insufficient time in which to make satisfactory judgements. There were many comments made which revealed considerable insight into the problems of non-native speakers of English while studying in Britain. The importance of academic, intellectual, personality, cultural and social factors was frequently mentioned alongside that of the linguistic ones. One comment struck the investigator as particularly perceptive and comprehensive and is reproduced below. The sum of the comments corroborates and complements much of what was said at a conference organised by BAAL and SELMOUS and recorded by Heaton and Cowie (eds, 1977), and could prove of value to other tutors and to teachers engaged in pre-sessional and in-session English courses.

Comment by a tutor on a lady student from Brazil (1592):

> The candidate's competent and intellectually sophisticated written work will clearly get her through the course, which is assessed by written term papers and by dissertation. If there were an oral assessment element based on active participation in seminar discussion she would be in peril - probably - though no more so than one or two of the native English speakers on the course. There has been almost no participation in the high-level cut and thrust of the seminars; how far this represents incompetence at that very exacting level of English usage (people alluding, trailling off in

tentative suggestion, interrupting, often fast or impassioned), how far a difficulty of cultural adaptation from (one suspects) a more passive and didactic university style, and how far an individual or cultural (woman: South American) tendency to passivity I really can't say - attempts at making things easier seem to induce tension so I've tolerated a totally silent presence in the seminars. This week for the first time a substantive intervention was made - a speech rather than discussion-exchange, though.

7.4.5 The important questions in the questionnaire were questions 1 and 2, and it was necessary to establish the reliability and validity of these particular measures before using them to establish relationships with pre-depature assessments. Information from the administration of the CPM was to be used when examining the external validity of the rating form, and correlations were used for examining the internal validity.

The first correlations to be computed were those between sections A and B of question 1 (R1A, R1B) and the grouped skills total of question 2 (RT). The correlations obtained were:

R1A with R1B	.78
R1A with RT	.85
R1B with RT	.78

(Table 7.37, Appendix IV)

All correlations between the different forms of rating were significant and high. Of particular interest was the correlation of .85 between responses at R1A - choice of four responses - and the grouped skills totals, where allowances were made for differences in performance in different skills. It has been argued that these were two approaches to assessing the same behaviour. The very high correlation is evidence of this and of the reliability of the two sections of the questionnaire. The slightly lower

correlations with responses to question R1B could be in part explained by the fact that the six alternative definitions do not so much refer to points on a continuum but to discrete, though related, definitions of performance. The correlations are nevertheless sufficiently high for an acceptable degree of reliability to be claimed.

In addition to the above, correlations between the four individual skills ratings were computed, which were all significant at the 1% level. (Table 7.38, Appendix IV.)

The correlations ranged from .59 to .72, showing a large degree of co-variance but sufficient difference to indicate that either different behaviours were being assessed or that different levels were being obtained in the different skills. The weakest relationship was listening with writing at .59, and the strongest was listening with speaking at .72. The individual skill ratings correlated much more strongly with the grouped skills total in every case at either .85, .86, or .87, from .14 to .19 above the highest skill to skill intercorrelation. This clear gap between intercorrelations of ratings between skills and of skills ratings with total can be considered evidence of a certain internal validity for this set of measures.

7.4.6 When the English Ability Rating forms had finally been analysed it was concluded that the required information had been collected on a sufficiently large sample with a broad range of backgrounds. It was further concluded that the rating form had proved a reliable measure with evidence of considerable validity.

It was hoped that this validity would be confirmed by results from the communicative proficiency measure and that the tutors' ratings would once again be shown to be strongly related to the pre-departure measures.

7.5 Communicative Proficiency Measure

The purpose of the measure was to assess directly the subjects' ability to communicate in real life situations (see section 4.6 above) in order to compare the conclusions reached with those of the tutors in their ratings. There was one major difference, however, between the measures. The assessment of performance on CPM was made according to a set of general criteria and to the investigator's notion of adequacy as applied to the notional 'average' non-native English speaking student. There was thus no account taken of other factors, such as course content, linguistic demands, or educational or cultural background. The content of the measure was presented in detail in paragraphs 7.1.4 to 7.1.4.4 above.

It was hoped to administer CPM to approximately 100 subjects - or approximately 10% of the total sample, as there was neither time nor finance to administer it to all the sample. It was hoped that the sample would prove fairly representative, although it was not possible to be so specific, since all participation by the subjects was voluntary and during term time.

7.5.1 The sample was established in two ways. The first was to make arrangements with special groups of overseas students whose tutors were sympathetic to the investigation and who were able and

agreeable to allotting a set time for the testing. Three such groups cooperated in the testing - 15 students in the Institute of Education (Overseas Education Unit) at Leeds University, 18 students from the Scottish Centre for Education Overseas, Moray House College of Education, and 21 in the Department of Administrative Studies, University of Manchester.

The second approach was to identify British Council offices and Colleges where members of the sample were concentrated and to send 'open invitations' to the students requesting them to cooperate. Approximately 200 subjects studying in Edinburgh and in six other centres were contacted in this manner. The catchment area covered just 30% of the total sample, and the 95 who finally took the measure amounted to 13% of the sample for whom English Ability Ratings were received. Exactly half of the CPM sample were studying in Scotland which compared with a total sample proportion of 7.7%. However, it was not felt that this was a significant aspect in the CPM sample's background. For logistic reasons the subjects studying in Scotland were bound to be exploited first. 54 (57%) of those taking CPM were members of special groups. That meant that of the 200 students contacted by individual letter only 41 responded - a rate of 1 in 5. (Table 7.39, Appendix IV.)

7.5.1.1 The characteristics of the background of the CPM sample are given briefly. The mean age was 31.8, almost two years more than the sample mean. 21 were female. The students came from 41 countries. Just over 60% of the sample came from countries in East and South Asia and South and Central America, and a further 19%

came from Middle Eastern countries. Sample members from African countries accounted for another 12% while the European countries were under-represented with a total of only 6 subjects. (Table 7.40) Of the countries sending the largest contingents, Mexico, Sudan, Brazil and Indonesia contributed larger proportions to the CPM sample. Thailand, Algeria, Turkey, Ethiopia and Japan, however, contributed very few.

Although larger proportions of the CPM sample came from Latin American and East Asian countries and smaller proportions from the Middle East and Europe, this was not considered to be a serious disadvantage.

Two thirds of the CPM sample had obtained first degrees or their equivalents in their home countries while almost a quarter had only completed secondary level education. This represented an increase proportionally over the main sample where 60% had completed tertiary and 20% secondary level education (see Table 7.6, Appendix IV.)

The distribution of subject areas studied was very different from that of the main sample because of the three special groups of students assessed. The Moray House and Leeds students were all following courses in the field of Education, including TEFL, and the Manchester group following a course in development administration were categorised in the Social Sciences (professional) group. (Table 7.41, Appendix IV.)

Since the investigator had no real control over participation or selection of the sample, nothing could be done to redress the balance. The CPM sample consequently lacked sufficient numbers

studying in more academic disciplines and particularly in Engineering and Technology.

A consequence of the presence of the three large groups was a significant change in the distribution of levels at which studies were being followed. Since all three groups were studying for Diplomas, the CPM sample was excessively weighted with subjects at this level. (Table 7.42, Appendix IV.)

The background characteristics of the CPM samples were reasonably similar to those of the whole sample, except in respect of subject areas and levels of study in Britain. The bias towards students following professional diploma courses in Education and Social Sciences had tbe beneficial effect of producing a sample of students from countries and in disciplines where proficiency in English had long been seen as an important consideration and indeed problem area.

7.5.1.2 37 members of the sample had taken EPTB before leaving their home country and 58 BCSA. This represented a heavier weighting towards subjects assessed by BCSA than in the whole sample - 61% compared with 56.5%. The EPTB group appeared to have slightly lower English proficiency than the whole EPTB sub-sample when distributions of Part 1 total scores and means were compared, and the BCSA group slightly better proficiency than the whole BCSA sub-sample. (Table 7.43, Appendix IV.)

70 members of the sample had taken an oral test, including 23 from the EPTB group, and 62 a writing test, including 16 from the EPTB

group. The means obtained were low B+ (3.4) for oral and B (3.1) for writing. (Table 7.44) There was no significant difference in performance between the two groups.

Almost three quarters of the CPM sample (70) had remedial English classes on arrival in Britain, compared with 63% of the whole sample. The average length of the tuition was 8.5 weeks, as for the whole sample. Only 1 subject had had more than 12 weeks tuition. (Table 7.45, Appendix IV.)

7.5.1.3 The general background profile of the CPM sample was found to differ from that of the whole sample principally in respect of subject areas and levels of study in Britain as well as being drawn in greater proportions from Asia, S America and Africa. Subjects were also on average slightly older and the proportion of female subjects was higher. No significant difference in pre-departure English proficiency was established, although there was evidence that the EPTB subgroup tended to slightly lower and the BCSA to slightly higher proficiency than the respective EPTB and BCSA main sub-samples.

7.5.2 The measure was administered as envisaged in paragraph 7.1.1 above, before and after the Easter vacation, mainly in the final week of term in March and in the first week of term in April. These were felt to be the times when students were most likely to find some time and when rooms were available. Since term dates differed from institution to institution the period over which the testing was done was quite extensive. The timetable finally agreed was:

March: Edinburgh, Leeds, Imperial College London, London School of Economics, and Southampton

April: Aberdeen, Edinburgh, Moray House, Glasgow

May: Manchester

Special groups are underlined. Administrations were carried out by the investigator only, although a second assessor was present for many of the interviews in Manchester. In general, insufficient time was available as students, who had been warned that the process would take up to 1½ hours, could either not wait beyond that time and so had to miss the interview, or could not even stay more than 1 hour which occasionally resulted in part of the reading or writing being left unfinished. Another factor was that many interviews continued beyond the planned 20 minutes or because interviewees did not wish to stop the conversation!

7.5.3 For most subjects the reading test was tackled first. Only one member of the CPM sample failed to attempt any of the cloze passages. Six subjects attempted two passages only in the time allowed, while twenty four subjects did not have time to attempt a fourth passage. Two thirds of the sample, however, attempted all four passages.

The exact word scoring method was used and so the means obtained were lower than if the acceptable word method had been used. Two of the passages, CC and MC, were of moderate difficulty with facility of 46% and 47%, while the other two were of greater difficulty with facility of 38%. Three of the tests discriminated moderately well, while test MC had a much greater discrimination - SD of 5.3 with 24 items, or 22%. (See Table 7.46, Appendix IV.) Test MC was

the modified cloze test in which only syntactic words had been deleted. With a reliability of .79 it was the most satisfactory sub-test of the four. It was also the test which fewest subjects attempted, partly no doubt because it was the fourth test to appear, and partly because the subject matter probably had little interest for the majority of the sample whose subject areas were education, administration and social sciences and engineering.

Total cloze scores were computed but no allowances were made for those who had not attempted all of the tests. The highest total score obtained was 85 out of a maximum of 109. Students did in fact complete the tests in any order, as requested, and some students obtained good scores in the two or three tests that they were able to attempt. The mean total score (TC) was 42 (38.5%). This relatively low percentage was accounted for by the fact that a third of the sample obtained zeros for one of the subtests and this was included in the total score. Zeros were not included in the computing of means for the individual cloze tests. In addition, the exact word scoring method kept the individual test scores low.

Reliability for the total score was satisfactory at .89 (Kudor Richardson 21 formula) as was to be expected from the length of the test and the relatively high standard deviation. Reliability for MC, at .79, was also considered satisfactory in view of its being the shortest subtest with only 24 items. Reliabilities for AC and CC were probably the best that could be hoped for from such short tests, while the reliability of test DC was low at .58. (Table 7.46, Appendix IV.)

The distributions of the scores on the four cloze subtests and the distribution for the total scores are given in figure 7.47 and figure 7.48 below. For the most part the distributions were slightly positively skewed, and tests CC and MC had the more evenly spread distributions.

7.5.3.1 The four cloze passages had been chosen in order to obtain a wide sampling of styles and subject matter. In addition one passage was a modified cloze passage with only syntactic words deleted in order to give the subjects some specifically syntactic problems. The results were examined to determine to what extent the choice of the different passages was valid.

Intercorrelations between the different cloze tests and total score were computed. It was assumed that correlations would be fairly high and positive. If one cloze test correlated with another at .8 or above it was felt that each would be testing much the same skills and therefore one might be redundant. Further, if MC25 was different in respect of topic and types of word deleted it was expected that correlations with other cloze tests should be lower than correlations between tests AC and CC for example.

All the correlations were positive and highly significant. (Table 7.49, Appendix IV.)

It was noted that correlations between tests AC, CC and DC (the pure cloze tests) ranged from .45 to .57 (rounded to two significant figures). From these it was concluded that each cloze passage was contributing something unique to the total. The correlations were

Figure 7.47 Distribution of scores on cloze tests (CPM)

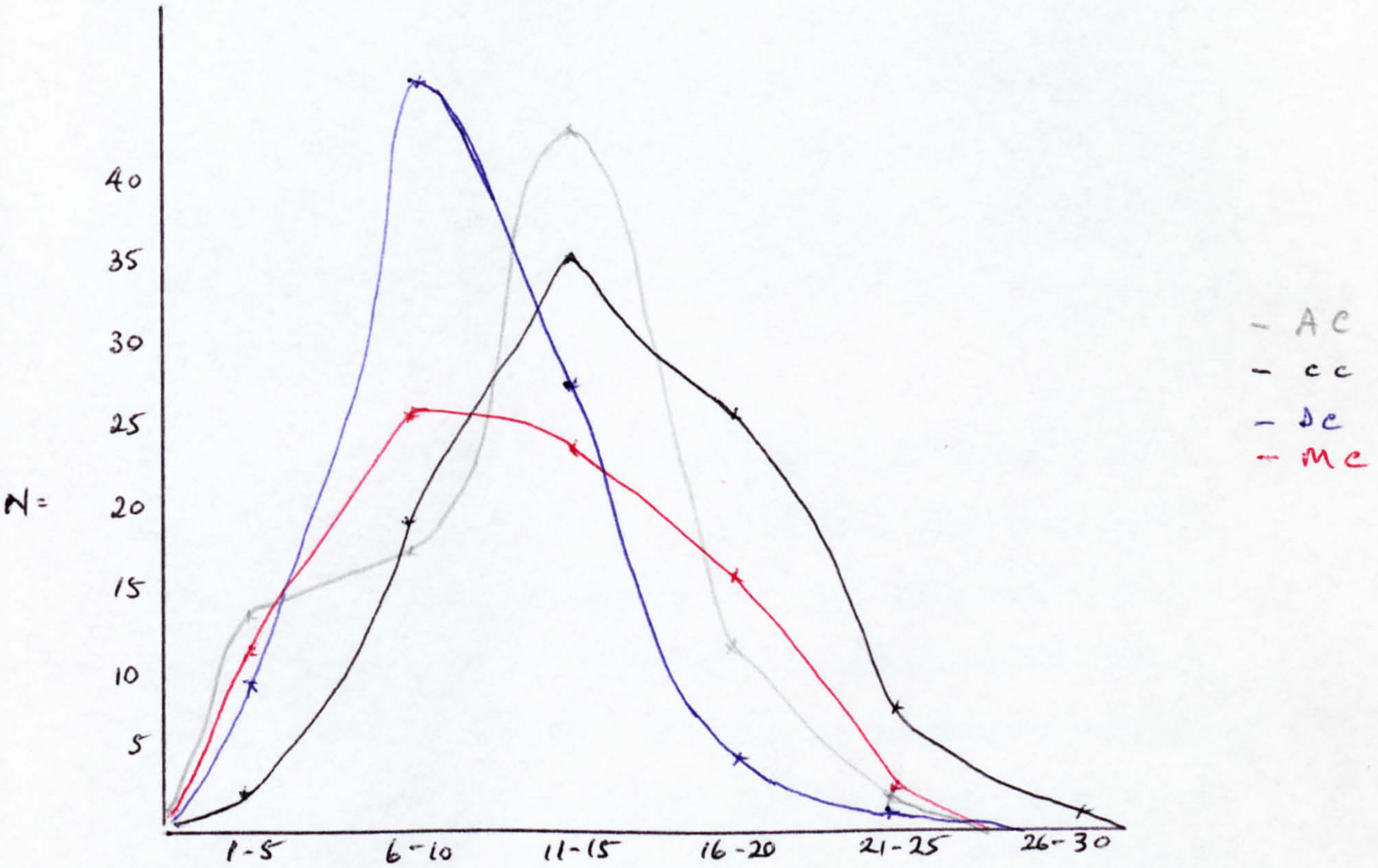
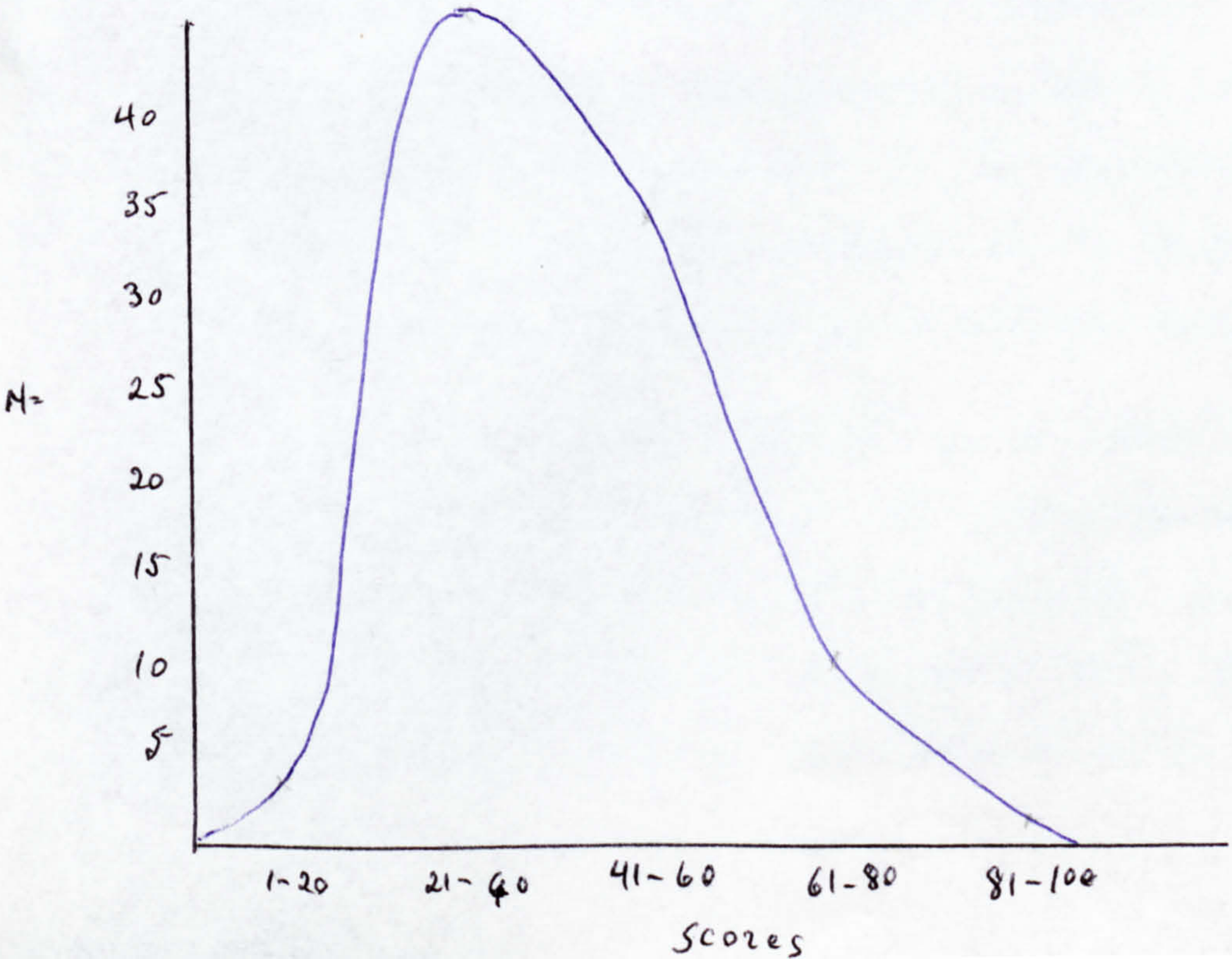


Figure 7.48 Distribution of total cloze test scores (CPM)



in fact lower than expected, and MC had low correlations with tests CC and DC as expected, at .36 and .34. It was, however, impossible to be certain exactly what the reasons were. The proposition that the differing subject matter meant real differences could well have been one explanation. The emphasis on wholly syntactic deletions would seem to be confirmed by the low correlations obtained by MC with two of the other tests. But why did MC correlate higher at .56 with AC?

The three pure cloze tests were analysed with respect to the function of the deleted words. It was noted that 16 of the 30 deletions in both AC and CC were syntactic words, whereas only 8 out of the 25 in test DC were. This could well account for the greater difficulty and lower correlations obtained between DC and the other tests. Its lowest correlation (.34) was with MC, the modified cloze test. However test DC was the least reliable of the subtests. The nature of the deletions may well have held the key, since on further analysis it was estimated that the syntactic deletions in test MC carried the lightest lexical load, and that the syntactic deletions in test AC carried a lighter lexical load than the syntactic deletions in test CC. The descending order of magnitude of test correlations with MC was AC, CC, DC in parallel with the increasing semantic load of the syntactic deletions.

Test correlations with the Total score were not unduly high considering that from one quarter to one third of the total score was contributed by the individual test in question. Although these test/total correlations ranged from as low as .62 to .78 they were from .1 to .21 higher than the inter-test correlations. This was

considered further evidence of the uniqueness of each sub-test.

7.5.3.2 It was concluded that the reading test as administered was an overall reliable test with acceptable discrimination that tested the reader's ability over a range of styles and likely academic subject matter. It remained to be seen whether the test appeared to be measuring proficiency.

7.5.4 In most cases the writing test was completed immediately after the reading. However, five subjects did not complete it. Where time permitted it was ascertained that candidates knew what was required of them. Although the topic of the writing was deliberately broad, it seemed to present very little or no problem for the candidates. In a few cases they were allowed to write on a topic of their own choosing. Most wrote four or five paragraphs as requested, and most finished within the time limit. The majority of essays were between 200 and 400 words long, with some (weaker) shorter ones and a few longer contributions.

The essays were assessed by the investigator according to the criteria set forth in paragraph 7.1.4.2 above. Scripts were marked first in small groups and levels assigned. When the whole sample had been tested, they were all reassessed and where necessary, adjustments made. The danger of unreliability in this method is readily acknowledged, but the stable factors were the definitions of the 6 levels, and the single assessor with his view of notional adequacy in English proficiency for non-native speakers. The full range of the scale was used, and it was felt that there was little

doubt about the writing awarded level 1. It was, however, sometimes difficult to distinguish between levels 2 and 3, or 3 and 4. The relationship between the number of inaccuracies and the degree of hindrance to the communciation was not always clear, for example.

7.5.4.1 Distribution of the levels was quite even with approximately three quarters of the sample each being awarded levels 3, 4 and 5. Table 7.50 below summarises the distribution for the writing test.

Only 5% were considered to have written inadequate English for academic writing, but a further 27.8% were adjudged to have produced writing that was barely adequate. Two thirds of the sample were adjudged to have written English which communicated the message adequately with little or no impeding of its comprehension.

Table 7.50 Distribution of levels for writing

Level		Frequency	%
1.	No communication	1	1.1
2.	Inadequate	4	4.4
3.	Serious weaknesses	25	27.8
4.	Some weakness but adequate	22	24.4
5.	Few weaknesses, very competent	23	25.6
6.	Close to N-S proficiency	15	16.7
	Total	90	100.0
	Missing	5	

7.5.4.2 The main problems facing the assessor were deciding to what extent the inaccuracies or inappropriacies of style, lexis and syntax impeded the communication and at what point these deviations

became unacceptable for a piece of academic writing. A further point which was difficult to resolve was determining whether unclear expression was due to inadequate control of the linguistic elements required or to poor organised ideas or to lack of them.

Eleven samples of the writing are contained in Appendix IV, Section 8, two from each level from 2 to 6 and the lone piece awarded level 1. The two samples of level 6 performance represent two different types of writing and the two major types of style encountered throughout the exercise. Subject 1961, a teacher in her mid twenties from Mexico, wrote in a very personal style and succeeded in conveying her mounting distress as the year progressed. The reader is not concerned whether this writer is a native speaker of English or not. Subject 320, a 39 year old lecturer from Bangladesh, wrote in a much less personal style. There are clues in the text which indicate that the author is not from a British cultural background, and the content is more of a political polemic. Yet, within the limitations of the content it is clear that the author has a near native proficiency in writing the language.

The samples of level 5 writing are by a young male teacher from the Solomon Islands (845) and a 38 year old lady teacher from Botswana (1315). The points they make are readily understandable and the personal opinions of the writers are well expressed. Yet each of them produced about 3 sentences in which there were either minor faults in grammar or cohesion which were more deviant from the norm than could reasonably be expected of a writer with near native command of English. The sentence (para 3 of essay 845)

> 'when September finally arrived and I was moved over to Leeds, my new residence shown and had to get down to some solid work, my attitude of Britain changed.'

is a valid attempt at a complex sentence, but the inappropriate switch from passive to active mood and the incorrect use of 'of' would not be expected of level 6 writers. The overall meaning of the writing and even the immediate meaning of the sentence were not impaired, and since such faults occurred infrequently the writing was considered entirely adequate. The sentence, 'there (is) nothing which I enjoyed most like schools observation' is a further example from para 1 of subject 1315.

In level 4 writing, as exemplified by no 25, a 32 year old scientist from Indonesia, and by no 584, a 37 year old agriculturalist from Cyprus, complex sentences such as those above tended to be more frequent and lower level mistakes began to appear. The following sentence by 25 illustrates the first point:

> 'If I mentioned about disadvantages I meant that being a married man it is not easy to be away from my family.' (para 2)

Transfer from the mother tongue is also evident in the use of 'mentioned about' and in the consistent omission of the definite article by subject 584, as in 'Population of Cyprus is about 600,000'. Subject 25 wrote in a personal style and his points were clear, while 584 wrote on the 'Economy of Cyprus' in a well organised and formal way but with a number of such type of errors.

The sentence

> 'Animal production in Cyprus is not a source of income which attribute to economy of Cyprus' (para 4)

is one which makes the reader pause briefly to substitute the correct lexical and syntactic forms and which typifies the lower limit of level 4 writing.

Level 3 writing, which is deemed barely adequate, tends to contain more seriously defective complex sentences than the one quoted above, and more of them. The reader is constantly having to pause to reinterpret. The following is an example from the essay of no 48, a 28 year old engineer from the Sudan

> 'In Scotland really I begin to understand something about the people of Scotland totally different from what I was first impressed with.' (para 2)

It was observed that at this barely adequate level of writing some sentences were difficult to understand because the basic idea contained was not clear and often not amplified in the adjacent text. Was the idea unclear because of linguistic problems or was the language unintelligible because of lack of ideas? The following sentence to illustrate this was written by subject 842, a 29 year old administrator from Mexico:

> 'So as to succeed in my studies I had to deal not only to my subjects but also as a human being who lives surrounded by other human beings.'

At this level of writing the rate of syntactic and lexical errors begins to rise. 842 only wrote 150 words but at least 10 unacceptable features were observed.

Level 2 was considered to be inadequate for academic writing and was marked by an increase in the rate of unacceptable expressions. Written pieces at this level also tended to be shorter. Subject 300, a 27 year old administrator from Peru, wrote only 110 words in the course of which at least 15 errors were observed. At this level of writing basic mistakes included errors in verb forms, spelling and inability to use connectors.

> 'Another important aspect of my experience in Britain was the way of life and the way how culd I adapted to it.' (para 2)

A 24 year old female teacher from Bahrein wrote a long (250 words) piece, but it was repetitious and the message was often quite unclear. The final sentence sums up the whole piece and leaves the reader wondering!

> 'Actually, writing a report for myself about this interesting school life was so beautiful, it really encourages me to take these reports to my country next year and work on it and do similar thing in my school if I could.'

In the opinion of the investigator the writer could well develop competent proficiency in English given time, but the cultural, professional and linguistic shocks that such a student must inevitably have suffered would take more than six months to overcome.

The one sample of a piece of writing which did not communicate was provided by subject 812, a 27 year old Jordanian with only secondary school background following a statistics course in a college of further education. 88 words were written in about 10 minutes, with all but 10 of the words monosyllabic. The theme sentence, possibly the most comprehensible, was 'Two things which I saw them very interested.'

7.5.5 In most cases the interview was conducted last, as planned. However, because of timing problems only 85 were able to attend the interview. Some time was spent in warming up and attempting to create a relaxed atmosphere. In a large number of cases this was not difficult since many of the students had volunteered presumably because they felt confident or were curious. There was

another group that never really relaxed. They were aware of their shortcomings in English and had attended the test, one suspects, in the hope that some remedy might have been offered!

It was not always easy to terminate the interview - a fact that had not been anticpated. As outlined in para 7.1.4.3 above, the topics of discussion were largely personal to the interviewee, and it became clear to the investigator as time passed, that for many of the subjects, the test must have offered the first opportunity for a long time for them to talk with a 'native' at length about their situation, their successes and their problems. In some cases certain phases of the interview, eg professional work, English language schools, yielded interesting and profitable discussion, sufficient to form judgements on, so that other phases had to be ignored because of lack of time.

The reliability of the procedure was suspect, but precautions had been taken within the severe restraints on time and finance that existed. The phases of the interview had been carefully worked out and adhered to as closely as possible. The order in which the phases were treated often varied according to the way the interaction developed. But the criteria used were the same throughout. The interviewer was the investigator and in this way inconsistencies in interviewing techniques and assessment were minimised. (During the final round of interview, at Manchester, Mr Ian Mackenzie attended and assessed the interviews according to the criteria and agreed with the investigator's ratings in 75% of the cases. In the remainder there was never more than one level's difference.)

The question of tape-recording was considered but in fact not adopted. It was felt that this would be a further imposition on volunteers who were already giving a lot of their time at an inconvenient period in the year. In addition many candidates were unused to being recorded and might even suspect that their 'bad' English might be broadcast more widely. An attempt was made with the Moray House group. Five subjects were asked if they would be willing to be recorded. Two objected, one spoke too softly to he heard clearly on the tape, and two samples remained. Carrying a tape-recorder from centre to centre and setting it up represented further work which the lone interviewer would have to undertake. The idea of recording was thus abandoned.

7.5.5.1 Ten students (11.8%) were found to be inadequate in their oral communication whereas almost three quarters (62 out of 85) were considered to be adequate or more than adequate. The distribution curve was nearer the normal curve than for the writing test, and the mode was at category 4 rather than at 3. Table 7.51 below summarises the distribution of levels awarded.

Table 7.51 Distribution of levels allotted for oral

Level		Freq	Freq %
1.	No communication	1	1.2
2.	Inadequate	9	10.6
3.	Serious weaknesses	13	15.3
4.	Some weaknesses but adequate	25	29.4
5.	Few weaknesses, very competent	22	25.9
6.	Close to N-S proficiency	15	17.6
	Total	85	100.0
	Missing	10	

One subject was awarded level 1. He was the only student to attend the administration of the test in Southampton. However, communication was not possible without lengthy pauses, breakdowns, repetitions and recourse to pencil and paper. The subject was studying for a Master's degree in Engineering. Attempts were made to get him to talk about his project. This proved to be (to the investigator's ears) /iɵtɵn dʌm/. When asked to write this it turned out to be 'earthen dam'. The candidate could not explain the principles of the dam even though he was asked to make a drawing. The investigator had no hesitation in awarding level 1, while nevertheless feeling that the student was in a near impossible position.

7.5.5.2 It is not possible to discuss the assessment of the interviews in detail in the absence of either transcripts or tape-recordings. However, brief notes were made for all candidates and certain features can be highlighted. The definitions given in para 7.1.4.3 above provided accurate general descriptions of the different kind of performances observed. Subjects awarded level 6 were generally very much at ease, fluent, and often had a good social manner. Good professional or personal discussions were often conducted. In general there was a full range of topics discussed, subjects exhibited mastery of a range of vocabulary and syntax, and in many cases had strong British accents. However, most did have some minor deviations from the norm in accent but were generally consistent. Five of the fifteen at level 6 had lived or worked previously in native English-speaking countries and this factor no doubt played a large part in their mastery.

During the interviewing the question 'what is being assessed - the person or the language?' kept recurring in the investigator's mind. There was no doubt that the personality and manner of the subjects did play a part. For example, a number of candidates obtaining level 5 grades were more hesitant or reticent than those obtaining level 6. They may not have been able to show the interviewer their full range of ability as a result. There were also male subjects at this level who were so voluble and almost uncontrollable that although their English was very fluent and accurate, their communication tended to be too one way and giving a 'steam-roller' effect! At this level L1 interference was often more marked in pronunciation, and difficulties were sometimes encountered in sytactically complex sentences. Communication was not impaired in any way, however, except occasionally by the personal qualities just discussed.

Level 4 subjects were still deemed to be communicating adequately for their academic purposes. The main features of performance at this level were an increase in the number, and sometimes length, of hesitations and searching for words. More cases of tentativeness, possibly lack of confidence, were observed. Limitations in the range of vocabulary and syntactic repertoire were also noticed. More regular and systematic pronunciation deviations were also noted, eg the omission of final consonants, deviant stress and intonation patterns. Systematic deviations from the syntactic norms were also occasionally noted, eg omission of articles. For the most part these deviations were sufficiently infrequent - or systematic - as to cause little disruption to the

conversation. 73% of those interviewed were allotted to level 4 or above.

Below level 4, however, significant changes occurred. Delivery was often slower, to allow time for thought or to concentrate on producing correct forms, and hesitations more frequent. The complexity of the sentences used was also reduced and many conversations were conducted mainly in simple sentences by the candidates. At this level candidates often experienced difficulty in understanding questions and remarks by the interviewer, and repetitions were requested. Discussion of English learning background often revealed a severe lack, frequently with little or no English having been learned at school.

A particular feature of level 2 and also level 3 subjects was a marked improvement in proficiency when subjects were talking about their own special field of work or study. This proficiency did not extend to more general and social interaction, which caused the investigator to award the lower level. A particularly interesting case was a subject (no 975) from France undertaking research in the biological sciences. He had brought his family from France and was living in the country outside Aberdeen. He had studied German as a foreign language at school, and certain German features came through into his English. His wife negotiated in English for everyday requirements, and he had refused an offer of 3 weeks' pre-sessional English!

It was noted that his vocabulary seemed reasonably extensive but he made frequent pauses while searching for what to say or how to

phrase his next contribution. His pronunciation had serious defects with French vowel quality and syllable-timed rhythm. However, when asked about his particular research project, the relationship of water with plants, he changed completely and was able to explain very clearly what he knew. The performance at that stage of the interview was bordering level 3 and 4 standard. But the overall impression was of an inadequate command of English for social and academic purposes. This was a case of a student who had tuned his English proficiency very finely. His wife took care of 'social English', and his command of his subject and terminology was such that he could discuss his immediate research interests to the satisfaction of his supervisor, who had rated his English as adequate with improvement desirable, and who had noted considerable improvement since his beginning his course.

Appendix IV, section 9, contains fuller details of the content of two interviews which resulted in the award of levels 5 and 2.

7.5.6 Once the three measures had been scored each subject had a profile of results consisting of one score (the cloze tests) and two levels (the writing and the interview levels). There was no overall score or level. Indeed any overall score would have obscured any differences there may have been between performance on the different tests. Pearson correlation coefficients were computed to establish whether any of the individual tasks were giving the same information. In addition, the writing and speaking grades were averaged to give a total productive skills grade. Correlations obtained were

Cloze with writing	.51,	
with speaking	.49,	with combined skills .55
Writing with speaking	.75,	with combined skills .93
Speaking		with combined skills .94

(Table 7.52, Appendix IV)

Two points emerged from the correlations. The reading test was clearly measuring different abilities. Secondly, the difference between performance on the writing and in the interview was not so great, although a correlation of .75 indicated sufficient variance to make the two tasks worth performing.

7.5.6.1 The three sets of results were next converted to parallel categories in order that some decisions could be reached about each individual student's adequacy, and in order to provide a basis for comparison not only within the CPM but also with the tutors' ratings. Categories parallel to those in section A of the tutors' rating questionnaire were chosen. Levels 1 and 2 for writing and speaking (interview) were considered to indicate that proficiency was not yet adequate, and recoded 1. Level 3 was considered to approximate to the category of 'only just adequate', and recoded 2. Levels 4 and 5 denoted adequacy and were recoded 3, while level 6 was clearly of an exceptionally high standard and was recoded 4 to correspond to the category 'completely adequate'.

The scores on the reading test were also converted to the same scale in an intuitive manner. Bearing in mind that only two thirds of the sample had attempted all four cloze passages, and bearing in mind that the exact word scoring method produced lower scores than the acceptable word scoring method would have done, the following bands of scores were proposed:

- under 27 (ie less than 25%)	- inadequate	1
- 27-36 (ie 25-35% approx)	- just adequate	2
- 37-60 (ie 35-55%)	- adequate	3
- over 60	- completely adequate	4

All scores and levels were recoded and are presented in Appendix IV, table 7.53. Distributions in the different categories by sub-test are set out in Table 7.54 below.

Table 7.54 Distributions of categories of adequacy, by CPM sub-tests

Category	Reading	Writing	Speaking
1. Inadequate	13 (13.8%)	5 (5.5%)	10 (11.8%)
2. Only just adequate	24 (25.5%)	25 (27.8%)	13 (15.3%)
3. Adequate	46 (49.0%)	45 (50.0%)	47 (55.3%)
4. Completely adequate	11 (11.7%)	15 (16.7%)	15 (17.6%)
Total	94 (100.0%)	90 (100.0%)	85 (100.0%)
Missing	1	5	10

The division of the reading scores into categories, while not established using a control group, yielded a distribution of similar proportions to the other two tests, although it appeared to be the most difficult of the tests. Since the correlations indicated that there was a large measure of variation in performance across the tests, further analyses were carried out. There were

31 cases who obtained the same level in all 3 tests

23 cases where reading was at the same level as either writing or speaking

19 cases where speaking and writing only were the same

7 cases who obtained different levels in all tests

15 cases where one or more test result was missing.

When differences in level occurred, these were normally of only one level. There were only six cases of levels differing by two

levels, reading being the test result concerned in four of these.

There were eighty cases for which both writing and speaking levels were recorded. In 50 of these there was no difference in category, as converted at Table 7.54 above.

7.5.6.2 Finally an attempt was made to attribute overall classifications to the subjects. This was done in two ways. Firstly the lowest category of the three awarded was taken as the final category. Cases with three different categories were ignored, and cases with assessments in only two tests but with the same levels were included. This analysis yielded the following final assessment:

Inadequate proficiency	8
Just adequate	18
Adequate	40
Completely adequate	9
	75
Incomplete	15

The second way was to make the final assessment correspond to the level attained in at least two tests, and where different levels were awarded for each test, eg 1, 2 or 3, to award the middle level of the three as the overall assessment. This method of classification was probably closer to the true picture of adequacy and yielded the following distribution:

Inadequate proficiency	5	(6.0%)
Just adequate	22	(26.2%)
Adequate	47	(56.0%)
Completely adequate	10	(11.9%)
	84	(100.0%)
Incomplete	11	

The principal disadvantage of this overall classification was the concealing of the differences in performance on the different skills in all but 31 cases ie in all but one third of the cases if the incomplete cases are included. Inter-test correlations highlighted the differences - particularly the difference in the nature of the cloze tests. The cloze test total consistently correlated lowest with the writing and interview and with the composite overall levels and the combined writing/interview level. There was a high correlation (.845) between the overall total and writing, which contrasted with the correlation between overall total and cloze (.663). Since no individual sub-test inter-correlations exceeded or even came close to sub-test/total correlations, this was taken as evidence of some degree of internal validity for the measures.

7.5.7 The communicative proficiency measure proved difficult to organise and administer with the limited resources available. However, the spirit of goodwill and cooperation on the part of students and staff enabled the investigator to administer the measures to 95 subjects throughout England and Scotland and to obtain complete results for 80, just over a tenth of the final sample. The test aroused a lot of interest among those taking it and only one case of hostility was encountered. The advisability of administering separate reading, writing and speaking tests was supported by the inter-test correlations, and performance on a wide range of linguistic skills was obtained in a relatively short time. The reading tests achieved a satisfactory level of reliability (.89) and every effort was made to reduce unreliability in the writing test and the interview through the structuring of the tasks and

the adherence to specific criteria when assessing the performance.

These features were considered to indicate a sufficient degree of internal validity for the results to be used to establish the nature of their relationship to both the tutors' ratings and the pre-departure assessments.

7.6 Conclusion

The second investigation was carried out with the aid of the English Ability Rating form, incorporating minor modifications as a result of experience in the first investigation, and the Communicative Proficiency Measure which had been trialled and finalised immediately after the first investigation. A sample of 925 students was chosen, and ratings obtained for 729. A sub-sample of 95 took the proficiency measures in addition, and 80 complete sets of results were obtained.

The sample differed a little in terms of country of origin and subjects of study from the first investigation sample. Pre-departure performance on EPTB and BCSA was not significantly different, although indications were that the EPTB sub-sample was of a slightly lower standard. Tutors' ratings again attained a high level of reliability, and the sample appeared to be a little more adequate in their proficiency than the first investigation sample. The CPM yielded information for which a certain degree of validity was claimed and at first sight appeared to support the tutors' judgements.

Comparison of the results of the ratings and the CPM, as well as examination of their relationships with pre-departure assessments and other relevant variables are reported in the next chapter.

CHAPTER 8 DISCUSSION OF RESULTS OF THE SECOND INVESTIGATION

8.0 Introduction

With the results of the tutor's ratings and the communicative proficiency measure collated three tasks remained. The first was to determine the validity of the information yielded by the criterion measures. The second was to establish the relationships between the pre-departure assessments and the criterion measures' assessments in the light of the hypotheses expressed in paragraphs 2.9.3 and 2.9.4. The third was to establish the relationships betwen the two sets of assessments for experimental groups affected by specific intervening variables as per the final hypothesis at paragraph 2.9.5. A further dimension to these second and third tasks was to monitor whether the patterns observed for the first investigation were maintained or varied in the second investigation sample.

8.1 The Criterion Measures

It has already been argued that evidence existed for a satisfactory degree of internal validity for both the tutors' ratings (see para 7.4.5 above) and the communicative proficiency measure (see para 7.5.7 above). It has yet to be established to what extent external or concurrent validity can be claimed for the measures.

The major problem in comparing the results obtained by the two instruments was the fact that the tutors were making general judgements while the CPM results were assessments of specific observed language performance in different modes, which might or might not have been representative of a student's regular performance. Neither assessment was being made along a continuous variable scale and consequently

comparison by correlation was not entirely suitable. However, a rectilinear relationship was anticipated.

It was decided nevertheless that firstly correlation coefficients should be computed between the ratings and the scores and secondly that the CPM overall results (see paragraph 7.5.6.2) should be compared with the tutors' overall ratings by means of tables.

8.1.1. The tutors' ratings as per their responses to question 1 sections A and B and their total skills ratings were correlated with each of the levels attained on the reading, interview and writing tests and with the composite interview/writing score and the overall category. Table 8.1 lists the correlations obtained. R1A = Rating question 1A, R1B = rating question 1B, RT = total skills rating, C = total cloze tests score, W = writing level, S = interview level, WS = composite writing and interview level, and O = CPM overall level. All correlations were beyond .01 significance. Full details are given in an expanded version of the Table in Appendix V.

Table 8.1 Correlation coefficients: tutors' ratings with CPM

		CPM				
		C	W	S	WS	O
Tutors	R1A	.411	.543	.531	.57	.521
	R1B	.488	.598	.606	.624	.612
	RT	.511	.656	.610	.656	.576

It was noted that, given the different nature of the assessment scales, the relationship between the two sets of measures was

positive and very strong. It was noted, however, that the cloze reading tests correlated consistently lower with the tutors' ratings than the oral and writing measures did. The difference between the correlations with cloze and the correlations with oral, writing and overall was only between 0.1 and 0.16. Had the difference been greater, say at around .3, it might have been inferred that performance on integrative cloze tests related much less to the general perception of language proficiency built up by a tutor over a few weeks or months than observed performance on language production tests did.

The relationship between tutors' assessment of writing and speaking abilities was of the same strength. Tutors' rating of writing correlated .639 with the CPM writing assessments and their rating of speaking .604 with the CPM oral. In view of the differences in the manners of the assessment these were considered adequate to indicate a strong and positive relationship.

8.1.2 Tutors' overall ratings, as exemplified by their responses to the categories in question 1A of the rating form, were compared with the investigator's overall assessment and attribution to identical categories. Tutors and the investigator assigned just over half the CPM sample to the same categories and differed by only one category in all but three of the remaining cases. The comparative distributions for overall adequacy according to tutors and the CPM assessment are presented in Table 8.2.

Table 8.2 Distribution of adequacy according to tutors and CPM

	CPM overall categories*					
Tutors' ratings	1	2	3	4	Total	%
a. Identical with CPM	0	7	30	6	43	53.8
b. One category higher than CPM	3	10	9	-	22	27.5
c. Two categories higher than CPM	2	-	-	-	2	2.5
d. One category lower than CPM	-	4	4	4	12	15.0
e. Two categories lower than CPM	-	-	1	-	1	1.2
Total	5	21	44	10	80	
%	6.2	26.2	55.0	12.5		100.0

* (1 = inadequate, 2 = only just adequate, 3 = adequate, 4 = completely adequate)

The table shows that 43 subjects were assigned to the same category by both tutors and the investigator. However, nine of the subjects assigned to category 3 by CPM were rated as 4 by the tutors (Column 3, row b), and 4 subjects assigned to category 4 by CPM were rated lower at 3 by the tutors (column 4, row d). These differences were largely academic since 3 represented adequate and 4 completely adequate. That distinction in terms of both real life for the students and value for the investigation was not very meaningful. The important distinctions concerned those deemed inadequate or only just adequate. Here there was less agreement between the two sets of measures. The 5 subjects evaluated as inadequate by CPM (column 1) were rated higher by the tutors, and the 5 subjects rated as inadequate by the tutors had all been evaluated by CPM as only just adequate (column 2, row d) or, in the case of one subject, as adequate (column 3, row e). The overall tendency of the discrepancies, however, was clear. CPM was evaluating the students'

proficiency at a slightly lower level than the tutors. It rated students lower than tutors in 24 cases and higher in 14.

For the purposes of the investigation these results were considered satisfactory. It was clear that the general relationship between the two measures was positive and very substantial. A larger and more representative sample would probably have produced a stronger relationship, but the principal factor was that the tutors and CPM were not necessarily defining adequacy in the same way. CPM made no allowance for the subject matter or nature of tuition of individual courses. It assumed a notional adequacy applicable to a theoretical or idealised student, based largely on the investigator's experience and intuition. Tutors, however, would have rated their students in the light of the particular circumstances of the course of study and may well have accepted lower standards of oral or written communication than the investigator. A further factor was that the overall categories derived from CPM were based on a profile of observed performances and were not strictly comparable with the tutors' ratings which were based on a cumulative general impression.

Those cases where tutors differed from CPM were examined further in an effort to identify any possible factors that might have accounted for the discrepancies. It was decided not to examine cases where the differences concerned only categories 3 and 4 (adequate and completely adequate). There were nine cases where students were rated adequate or only just adequate by CPM and one category lower by the tutors. 6 of the 9 were studying social

sciences (professional). Initially this high proportion apeared to lend support to the observation that students in this field, and in engineering, might need greater proficiency in English than is generally assumed (end of para 6.4.3.4 above). However, since six of the nine were following the same course in Manchester it was quite possible that the tutor was particularly strict in the standards he expected.

15 subjects evaluated by CPM as inadequate or only just adequate were rated higher by the tutors. 5 were studying science subjects, 3 education and 5 social sciences (professional). The two latter subject areas each accounted for 30% of the CPM sample and so their numbers were not disproportionate, although two of the social science students were also from the Manchester course. It was interesting that 5 of those rated higher were studying the pure sciences/maths. However, three of these were following a tertiary level course in statistics in Aberdeen. Again, the tutor may have simply been less aware of the English of his three students, or the study of statistics in that particular course may have required a comparatively low level of English proficiency.

Of the 24 discrepancies discussed above it was noted that 14 were involved in three particular courses - 8 were following the Development Administration diploma course in Manchester, 3 the Urban Design and Regional Planning diploma course in Edinburgh, and 3 the Statistics diploma course in Aberdeen. Many of the discrepances may, therefore, have been a feature of particular courses or even tutors. The main inferences to emerge from the examination of the discrepancies were, that the English needs of courses with large

numbers of non-native speakers of English could profitably be examined and pre-departure English assessment norms established, that professional diploma courses in the social sciences may well demand a higher level of English proficiency than generally assumed and that academic courses in maths and sciences may sometimes require a standard of proficiency that is a little lower than generally assumed.

8.1.3 Although the analysis so far was considered to have indicated sufficiently similar information deriving from the two assessments, a 2 x 2 contingency table was established. If the two measures were making the same fundamental distinction between adequate and inadequate English proficiency it was expected that the number of 'hits' ie those deemed adequate by both measures and those deemed inadequate by both measures, would be high. Table 8.3 presents the results. Column 1 and row 1 indicate inadequate English, and column 2 and row 2 denote adequate English, represented by categories 2, 3 and 4 in Table 8.2

Table 8.3 Contingency tables: tutors' ratings and CPM results

		CPM 1	CPM 2
Tutors	1.	0*	5
	2.	5	70*

* confirmed expectancies rate = 87.5%

According to this basic two way analysis the two measures agreed in 87% of the cases for which tutor ratings and CPM results were obtained.

8.1.4 The results of the comparison of information given by the two measures led to the conclusion that both were giving the same fundamental information and differing to an acceptable degree when the assessment categories and scales were refined. It was concluded that an acceptable degree of concurrent validity could be attributed to the two criterion measures.

8.2 Comparison of Pre-departure with Criterion Assessments

The same methods of comparison were used as for the first investigation. It was anticipated that the correlations would be low and similar to those of the first investigation. It was anticipated that the tables would yield to within 10% the same rate of hits and efficiency over chance. In addition to analysing the results of the total sample, the results of the CPM sample were further examined separately. As in the first investigation the comparisons were carried out in two sub-samples according to the pre-departure test taken.

8.2.1 Correlations between the total Part 1 EPTB scores and the tutors' ratings in question 1 and their rating of the skills in question 2 were computed. Correlations of the BCSA notional totals with the same tutors' ratings were also computed. Table 8.4 summarises these correlations.

Table 8.4 Correlations: pre-departure assessments with tutors' ratings

(a) EPTB Pt 1 with tutors' ratings (R1A)	.306 N = 311 S = 0.00
(b) EPTB Pt 1 with tutors' ratings (R1B)	.270 N = 311 S = 0.00

(c) EPTB Pt 1 with tutors' skills ratings (RT) — .357, N = 296, S = 0.00

(d) BCSA notional total with tutors' ratings (R1A) — .364, N = 368, S = 0.00

(e) BCSA notional total with tutors' ratings (R1B) — .343, N = 371, S = 0.00

(f) BCSA notional total with tutors' skills ratings (RT) — .359, N = 352, S = 0.00

The correlations were above .3 and highly significant, as for the first investigation. The correlation of EPTB Pt 1 with RT was identical to that of BCSA total with RT and the lowest correlations were those with the 6 point rating scale, R1B. The BCSA correlation at (f) was almost identical with that obtained for the first investigation sub-sample, while that of the EPTB sub-sample was higher at .357, compared with .319 in the previous investigation. In view of the lack of ideal conditions for correlations, as discussed in para 6.2.4, the correlations reported above were considered to be indicative of a strong and positive relationship between the pre-departure and tutors' assessments.

A check was carried out to establish whether correlating total with total yielded the strongest relationship, or whether a sub-test score on either EPTB or BCSA had correlated more highly with RT. It was found that no sub-test score on either measure correlated more highly than the total measure score. It was, however, noted that the grammar subtest in EPTB correlated almost as highly at .331, while the writing test grades for BCSA and EPTB sub-samples

correlated less than .01 lower than the total with RT. It could therefore be inferred that the discrete point grammar test and the writing assessment had the greatest potential as predictors among the various subtests.

The most striking correlation, however, was that of reading speed (EPTB test 5) with total ratings (RT) at .402, almost .05 higher than the total with total correlation. About 40% of the EPTB sub-sample took this test and so the result must be considered meaningful. Little work on the intrusive word technique has been done recently, but this relationship, if repeated elsewhere, seems to indicate that this kind of reading test may well possess a large measure of the construct of proficiency in another language and thus be a strong indicator of language proficiency. It is a test which is simple to devise, takes only a short time to administer and yields a large number of items with a high degree of reliability. (See Table 3.2 above)

Table 8.5 reports the correlations between sub-tests and tutors' total ratings. It was not easy to draw comparisons with the first investigation results, since the rating scale of question 2 of the English Ability Rating form had been modified for the second investigation and four labels were provided instead of three (see Figure 7.28 above). Indications are that the modification has strengthened the relationship between the measures, as all sub-test total correlations in the second investigation were at least .06 higher for the EPTB sub-sample and for listening and writing for the BCSA sub-sample. Table 8.6 reports the correlations between the pre-departure skills ratings and the tutors' ratings. (Fuller details are given in Tables 8.5.1 and 8.6.1 in Appendix V.)

Table 8.5 Correlations: EPTB and BCSA sub-tests with tutors' ratings (RT)

	EPTB sub-tests	r	N	S
1.	Phonemic discrimination	.115	263	0.03
2.	Intonation	.226	263	0.00
3.	Reading comprehension	.299	268	0.00
4.	Grammar	.331	268	0.00
5.	Reading speed (Part 2)	.402	127	0.00
6.	Oral/interview	.360	157	0.00
7.	Writing	.350	140	0.00
	BCSA sub-tests			
1.	Listening	.291	347	0.00
2.	Speaking	.287	347	0.00
3.	Reading	.272	331	0.00
4.	Writing	.357	331	0.00

Table 8.6 Correlations: EPTB/BCSA individual skills with tutors' ratings of individual skills

EPTB	r	N
Speaking with speaking	.361	169
Writing with writing	.332	140
BCSA		
Listening with listening	.285	366
Speaking with speaking	.333	367
Reading with reading	.232	343
Writing with writing	.344	332

It was noted that the speaking and writing correlations were of similar magnitude to the total with total (RT) correlations (see Table 8.4) as well as to the skill with total (RT) correlations, with the exception of BCSA speaking. Once again the explanation could have been that speaking and writing performance is more easily observed and so more confidently assessed than the passive skills. The grading and rating of reading and listening skills both with BCSA and by the tutors may well have been carried out with much less certainty.

8.2.1.1 Twenty-nine correlations of pre-departure totals and sub-tests with tutors' ratings of skills have been reported above. Only one, the EPTB phonemic discrimination with total ratings, was less than .22 and not highly significant. The others ranged from .22 to .40 and were all significant beyond .01. The most important correlations, totals with total, were above .35, and speaking with speaking and writing with writing correlations ranged from .33 to .36. The correlations were generally higher than those obtained in the first investigation and demonstrated clearly a very positive and substantial relationship between EPTB and BCSA assessments and tutors ratings.

8.2.2 The comparisons were next carried out by means of contingency tables - see paras 6.1.2 and 6.3 above. The basic distinction to be made was whether the English proficiency of the subjects was judged to be adequate or inadequate. The EPTB and BCSA assessments were grouped into categories on the basis of total Part 1 scores and notional total ratings as for the first investigation (see paras 6.3.1 and 6.3.2 above). These groupings are summarised as follows:

	EPTB Part 1	BCSA grades
1. Inadequate	less than 34.0	C+ or less
2. Just adequate (remedial tuition needed)	34.0 to 39.9	B
3. Adequate	40.0 to 45.9	(B+, A
4. Completely adequate	46.0 and above	(

Comparisons were made with the tutors' responses at question 1A and 1B and with the total ratings in question 2, grouped in categories as in para 7.4.2.1 above.

The contingency tables between the EPTB or BCSA and tutors' ratings are listed in Table 8.7. The cross tabulations on which they are based are in Tables 8.7.1 to 8.7.6 in Appendix V. The EPTB/BCSA distributions are expressed in the rows and the tutors' ratings distributions in the columns. Row 1 indicates inadequate English, and row 2 indicates just adequate, adequate or completely adequate as shown by EPTB or BCSA. Column 1 denotes tutors' ratings of inadequate in question 1A (R1A), ratings of inadequate and more proficiency essential in question 1B (R1B) and ratings of inadequate in question 2 (RT). Column 2 denotes adequate - the amalgamation of the remaining categories of R1A, R1B or RT as appropriate. Each cell contains the number of cases together with percentages in brackets. The percentage rate of confirmed expectations (hits) is also shown.

Table 8.7 Contingency Tables: EPTB/BCSA and tutors' ratings

(a) EPTB and Ratings (R1A)

		R1A 1	R1A 2
EPTB	1.	3 (1.0)	49 (15.8)
	2.	6 (1.9)	253 (81.3)

Confirmed expectations = 82.3%
Source: Chi square = 44.04
df = 9, S = 0.00

(b) BCSA and Ratings (R1A)

BCSA	R1A 1	R1A 2
1.	3 (0.8)	77 (20.9)
2.	4 (1.1)	284 (77.2)

Confirmed expectations = 78%
Source: Chi square = 50.74
df = 6, S = 0.00

(c) EPTB and Ratings (R1B)

EPTB	R1B 1	R1B 2
1.	5 (1.6)	47 (15.1)
2.	20 (6.4)	239 (76.8)

Confirmed expectations = 78.4%
Source: Chi square = 39.49
df = 12, S = 0.00

(d) BCSA and Ratings (R1B)

BCSA	R1B 1	R1B 2
1.	6 (1.6)	74 (20.0)
2.	13 (3.5)	278 (74.9)

Confirmed expectations = 76.5%
Source: Chi square = 48.75
df = 10, S = 0.00

(e) EPTB and Ratings (RT)

		RT 1	RT 2
EPTB	1.	1 (0.3)	49 (16.9)
	2.	5 (1.7)	236 (81.1)

Confirmed expectations = 81.4%
Source: Chi square = 38.84
df = 9, S = 0.00

(f) BCSA and Ratings (RT)

		RT 1	RT 2
BCSA	1.	1 (0.3)	78 (22.1)
	2.	1 (0.3)	272 (77.3)

Confirmed expectations = 77.6%
Source: Chi square = 31.79
df = 6, S = 0.00

It was noted that the percentages of confirmed expectations were higher with the EPTB sub-sample and that in both sub-samples comparison with both R1A and total skills ratings yielded very similar percentages. With a success rate of over 80% for EPTB and just under 80% for BCSA it was felt that the strength of the relationship between pre-departure and tutors' assessments was clearly demonstrated. The increase in efficiency over chance ranged from 53% (BCSA and R1B) to 64.8% (EPTB and RT).

The success rate was about 5% lower than for the first investigation for both sub-samples, but the better results obtained for EPTB were repeated in this investigation, as was the slightly lower success rate when R1B was used.

The tables presented in Table 8.7 were obtained from cross-tabulations involving from 12 to 20 cells. Chi squares were obtained for all these and in every case were found to be highly significant.

One feature of the tables was the rate of misses where EPTB or BCSA predicted adequate English and where tutors found the subjects to be inadequate. When R1A and R1B were used the rate was never above 2% either for EPTB or for BCSA. These were lower rates than for the first investigation. The higher rate of such misses with R1B may be attributed to the fact that the investigator classified "more proficiency essential" as being inadequate. In many of the cases tutors must have considered these cases to be borderline.

8.2.3 As for the first investigation the contingency tables confirmed the strength of the relationship of both EPTB and BCSA with the tutors' ratings. Correlations obtained at above .35 were considered highly satisfactory, and the strength of the relationships between some of the sub-tests - EPTB grammar and reading speed plus the assessment of oral and written tasks - with tutors' ratings was also noted. Since it had already been established (para 8.1.4 above) that both the tutors' ratings and CPM were providing much the same information, it was expected that the positive and substantial relationship would be confirmed by the CPM sub-sample.

8.2.4 The CPM sample was much smaller, and consequently it was anticipated that correlations might be lower and in some cases non-significant, as the sample had to be again divided into EPTB and BCSA sub-samples. (Breakdowns of background details are given in

Tables 8.8 to 8.11 in Appendix V.) Nevertheless performance on the pre-departure assessments was compared with the overall and detailed assessments obtained by the CPM, and correlations and contingency tables were established.

8.2.4.1 The EPTB group of the CPM sample for whom results were obtained consisted of only 34 subjects. Nearly half were following social science (professional) diploma courses and a third education diploma courses. Just over a tenth were studying engineering. Three quarters had obtained between 34.0 and 39.9 on the EPTB pre-departure test and the remainder were divided equally between those who had been judged inadequate (scores of less than 34.0) and those with scores of 40 to 45.9. The EPTB mean score was 37.6 - lower than that of the whole sample. Most had had remedial tuition. Because of the bunching of the EPTB scores and subject areas and levels of study, it was anticipated that correlations might be low.

Thirty correlations (see Table 8.12 in Appendix V) between EPTB sub-tests and totals and CPM sub-tests and overall levels were computed - Tests 1-4, Pt 1 Total and Pt 2 with CPM cloze, oral interview, writing, combined oral/writing, and overall category. The cloze correlations were calculated with raw scores, the rest with categories, 1-4 or 1-6 as appropriate. 16 of the correlations were significant beyond the 5% level. None of the correlations with tests 1 and 2 (phonemic discrimination and intonation) was significant. Test 3 (reading comprehension - modified cloze) did not correlate significantly with the cloze tests but correlated at .33 with the writing. Test 4 (grammar) correlated .32 with CPM

writing, .42 with the interview and .38 with cloze. Part 1 total correlated positively with all CPM tests and combinations from .314 to .336, except with the cloze reading - for which the correlation was .436. These were considered satisfactory, particularly as they were of the same order as the Part 1 total/tutors' rating total correlations, even though the group consisted of only 10% of the EPTB sample. The correlation with the cloze test was a full .1 higher than the other correlations. This could well have been due to its greater similarity in nature with EPTB than the other components of CPM. Further correlations between the oral ratings and CPM sample were obtained, since two thirds of the CPM group had been given a pre-departure interview rating as well. All the correlations were significant beyond the 1% level and higher - from .52 (with cloze) to .64 (speaking and writing combined).

The most striking correlations were for EPTB Part 2 (reading speed) and CPM. Only 13 subjects had taken this test and CPM. Nevertheless all five correlations were significant beyond the 1% level and the highest reported so far. They are set out below:

EPTB Pt 2			
with CPM overall	.692	(N = 12)	P = .006
with CPM interview	.741	(N = 12)	P = .003
with cloze total	.757	(N = 13)	P = .001
with CPM writing	.851	(N = 13)	P = .000
with speaking/writing combined	.857	(N = 12)	P = .000

The relationship indicated here was very substantial indeed, particularly with the writing performance and the combined writing and interview level. The consistently high correlations would tend to discount the idea of the nature of this small sub-sample being the explanation. It was noted in para 8.2.1 above that Part 2 correlated highest with the tutors' ratings. The tendency has

been repeated in the correlations with CPM.

8.2.4.2 Similar correlations were computed for the BCSA group of the CPM sample. This group numbered 56 and was more evenly spread across the subject areas, except that two fifths were following Education diploma courses and a quarter social science (professional) diploma courses. The group was on average 2 years younger than the EPTB group with a mean BCSA total grade of B+, as for the whole sample.

Twenty-five correlations (see Table 8.13 in Appendix V) between BCSA sub-tests and total and the CPM sub-tests and overall levels were computed. The cloze correlations were calculated using groupings of scores as in para 7.5.6.1 above. Since this scale was closer in nature to the A to E grading of BCSA it was more suitable than the raw scores for correlation purposes. Twenty-two of the correlations were significant beyond the 5% level. The BCSA listening grades were the only grades to produce non-significant correlations. BCSA speaking correlated from .41 to .68 with the CPM measures, reading from .34 to .55 and writing from .43 to .58. The notional totals correlated from .52 to .66. The general range of correlations was not very great and speaking provided the greatest range. No correlations approached the level of the EPTB Part 2 correlations with CPM.

The higher correlations for the BCSA sample were probably due to the larger numbers in the group, the more subjective nature of both sets of measures compared with EPTB and the fact that both sets of scales being correlated were limited in length.

8.2.4.3 The correlations between BCSA and CPM were considered highly satisfactory and indicative of a substantial and positive relationship between the pre-departure BCSA assessments and the CPM 'in study' assessment. The relationship between EPTB and CPM was not so strong. However, this was nevertheless considered to be satisfactory as the correlations with Pt 1 Total were all above .3, the sample was very small and restricted in pre-departure English proficiency, and the scales of the measures being compared were totally different - except in the case of raw cloze scores (CPM). In addition the Part 2 scores indicated the possibility of an even stronger relationship with 'in study' English proficiency.

8.2.4.4 The correlations between performance on EPTB or BCSA and CPM were for the most part greater than those obtained between BCSA or EPTB and tutors' ratings for the whole sample. Similar correlations for the CPM sample only were computed. Since the sample was much smaller, it was anticipated that the correlations might be lower.

EPTB Part 1 (CPM sample) total did not correlate significantly with any of the tutors' ratings. 30 correlations of EPTB sub-tests plus total with tutors' skill and total ratings were computed (see Table 8.14 in Appendix V). Of these only 6 were significant - modified cloze with reading (.49), grammar with writing (.3) and Part 2 with all except listening (range .69 to .86). The lack of significant correlations indicated that the relationship with tutors' ratings was not as strong as that with the Communicative Proficiency Measure, although the correlations obtained with Part 2 were of the same magnitude.

25 correlations between BCSA (CPM sample) and tutors' ratings were computed. (See Table 8.15 in Appendix V.) Of these only 10 were significant. No positive correlations were obtained with BCSA listening and reading. BCSA speaking correlated significantly with tutors' listening, speaking and total ratings from .38 to .28; writing .33 with tutors' ratings of writing, .3 with speaking and .29 with total; BCSA total correlated with all skills except writing and with tutors' total ratings at .297. This total/total correlation was lower than for the whole sample, and these results were therefore taken to indicate that the relationship between pre-departure assessments and tutors' ratings for the CPM sample was weaker than for the whole sample, as revealed by correlations. The relationship between EPTB/BCSA and CPM was seen to be a stronger one.

8.2.4.5 In order to confirm this strength contingency tables were also established. It was anticipated that students whose pre-departure English had been deemed inadequate would remain so and that the English of those who had been assessed as just adequate or better would prove to be at least just adequate. In Table 8.16 below row 1 indicates that the EPTB score was inadequate and row 2 that the EPTB score was just adequate or more. Column 1 represents inadequate proficiency during studies in Britain, and column 2 only just adequate or adequate, as determined first by the CPM overall rating and secondly by the combined oral/writing level obtained.

Table 8.16 Contingency tables - EPTB with CPM

CPM (overall)

EPTB	1.	2.
1.	0*	5
2.	2	28*

Expectations confirmed = 80%
Source: Chi square = 21.6
df = 6, P = .001

CPM (oral/writing)

EPTB	1.	2.
1.	0*	5
2.	2	27*

Expectations confirmed = 79.5%
Source: Chi square = 32.4
df = 16, P = .009

The confirmed expection rate at 80% was considered satisfactory in view of the size of the sample and was 60% more efficient than chance. This rate was comparable with the rate for the total sample and tutors' ratings, and a little higher than the rate for EPTB and tutors' ratings for the CPM sample, as in Table 8.17.

Table 8.17 Contingency tables - EPTB with tutors' ratings (CPM sample)

Ratings

EPTB	1.	2.
1.	0*	5
2.	3	24*

Expectations confirmed = 75%
Source: Chi square = 15.2
df = 6, P = .018

8.2.4.6 Tables established for the BCSA group (see Tables 8.18 and 8.19 below) produced the same rate of confirmed expectations with CPM and a higher rate for the tutors' ratings. All tables were drawn from cross-tabulations with significant chi squares.

Table 8.18 Contingency tables - BCSA with CPM

CPM (overall)

BCSA	1.	2.
1.	0*	7
2.	3	39*

Expectations confirmed = 79.5%
Source: Chi square = 16.3
df = 6, P = .012

CPM (oral/writing)

BCSA	1.	2.
1.	0*	6
2.	2	38*

Expectations confirmed = 82.5%
Source: Chi square = 28.6
df = 16, P = .026

Table 8.19 Contingency tables - BCSA with tutors' ratings (CPM sample)

Ratings

BCSA	1.	2.
1.	0*	9
2.	0	42*

Expectations confirmed = 82%
Source: Chi square = 17.1
df = 4, P = .002

8.2.5 Three conclusions were drawn from the above comparisons. Firstly, the relationship between EPTB/BCSA and the criterion measures was shown to be positive and substantial for the whole sample both by the correlation method and by the contingency method. Secondly, the relationship between the pre-departure measures and tutors' ratings as expressed by correlations seemed to be of equal strength for both sub-samples, although contingency tables showed a marginally higher rate of confirmed expectations for the EPTB sub-sample. The third conclusion was that the relationship between the CPM and BCSA, as revealed by the correlations, appeared stronger than between CPM and EPTB. This may have been in part due to the very small size of the EPTB group. But it was also noted that EPTB Part 2 correlated higher than any other individual measure with the CPM measures, and that this might be an indication of a potentially stronger relationship between CPM and EPTB.

8.3 Comparisons Affected by Different Variables

Comparisons were next made between pre-departure and criterion assessments in the light of certain intervening variables. The sub-samples were divided into different experimental groups and their performances compared. Relevant data are given in Appendix V, Tables 8.20 to 8.27.1. Comparisons were made with tutors' ratings only, since the CPM sample was too small.

8.3.1 The first experimental groups were established according to sex. Only one sixth of the sample was female. There was little difference in their English proficiency on departure from that of the male group. Distributions of grades for BCSA were almost

identical (see Table 8.20 in Appendix V), but for EPTB 42% of the female group obtained scores above the mean of 40 while only 31% of the male group achieved this. Background education was almost identical but there were differences in subject areas of study. Proportionately more women were studying in the fields of Education, English and Applied Linguistics, Arts and Medicine, while the proportion of men studying Engineering subjects was five times greater.

Means for tutors' ratings were almost identical for both groups. However, a larger proportion of the male group were adjudged to have made considerable improvement in their English (21.5% compared with 13.7%). Correlations of EPTB/BCSA with tutors' ratings were positive and significant, except for the EPTB/Rating 1A correlation for the female group. Correlations for the male group were of the order of the whole sample, whereas the positive correlations for the female group were a little higher, from .39 to .46. Contingency tables (see Table 8.21, Appendix V), revealed that the rate of confirmed expectations for EPTB was 85% (female group) and 81% (male group) and for BCSA 81% (female group) and 77% (male group). These were at approximately the same rate as for the whole sample, and with EPTB again maintaining a slightly higher rate than BCSA.

In conclusion it was observed that the relationship between pre-departure and criterion measures was positive and substantial. There seemed to be no significant difference between the two groups. The female group was smaller and showed a tendency towards a slightly stronger relationship, which could have been due to their following the more academic subjects.

8.3.2 The sample was next divided into one group that had followed Remedial English classes before embarking on their main course of study - referred to below as the 'plus remedial' group - and one group that had not - referred to below as the 'no remedial group'.

In general background the 'no remedial' group (N=269) had had more education before reaching Britain, 30% having obtained higher degrees as opposed to 14% of the 'plus remedial' group. More than a quarter of the no remedial group were studying for a research degree in Britain and 43% were on attachment or following diploma courses. For the 'plus remedial' group the proportions were one tenth and 60%. More of the students studying engineering took remedial English than of any other subject group. (Details are in Table 8.20, Appendix V.)

In the first investigation the difference in proficiency between the samples was examined (para 6.4.1.2). It was noted that there was over one SD's difference in pre-departure English performance and that there was still ½ to ⅔ SD difference between the two groups when the tutors' ratings had been collated. These differences were repeated in the second sample with a slight increase in difference after the tutors' ratings, as set out in Table 8.22 below.

Tutors noted that some or considerable improvement in English proficiency had been made in 62% of the cases in the 'no remedial' group and in 84% of the 'plus remedial' group. From these figures it can be inferred that the beneficial effect of the remedial courses was carried over into the main study period itself. On the other hand 20% of the 'no remedial' group had made no improvement

Table 8.22 Means and SD's for the remedial English groups

			Pre-departure	Rating (RT)
EPTB Pt 1	Total sub-sample	M	38.3	35.6
		SD	4.5	7.6
	No remedial group	M	40.6	38.3
	Plus remedial group	M	35.3	34.2
	Difference		over 1 SD	over ½ SD
BCSA total	Total sub-sample	M	13.3 (B+)	38.6
		SD	2.2	7.1
	No remedial group	M	15.1 (A-)	42.1
	Plus remedial group	M	12.3 (B)	36.3
	Difference		1¼ SD	over ¾ SD

because of their high initial proficiency. For the 'plus remedial' group the figure was only 8%. This relatively high figure could be accounted for by the fact that almost a third of the BCSA sample following remedial English had BCSA grades of B+ or more. Nevertheless, given the high improvement rate, there would appear to be advantages for students even with adequate pre-departure proficiency in following a short course of remedial English.

There was a marked difference in the two groups' relationships between pre-departure proficiency and 'in course' proficiency as shown by both correlations and contingency tables. Two correlations of .42 were obtained for EPTB Pt 1 total with tutors' ratings for the 'no remedial' group, while they were only .13 and .19 for the 'plus remedial' group. The rate of confirmed expectancies was 96% for the former and 75% for the latter (see Table 8.23, Appendix V). Correlations for the BCSA sub-sample were unexpected. They were positive and significant in the case of the 'plus remedial' and non-significant for the 'no remedial' - the reverse of the situation for the EPTB sub-sample. However the non-significant correlation

was due probably to the fact that that group all had scores of B+ or A and so were not distributed in a satisfactory manner for correlation. This was supported by the contingency tables where 97.5% of the expectancies were confirmed for the 'no remedial' group. Correlations and rate of confirmed expectations are summarised in Table 8.24 below.

Table 8.24 Correlations for the remedial groups

			Confirmed expectancies
EPTB Pt 1			
with tutors' ratings	(R1) - no remedial group	r = .42	96%
" " "	(RT) - " " "	r = .42*	
" " "	(R1) - plus remedial group	r = .13	75%
" " "	(RT) - " " "	r = .19	
BCSA			
	(R1) - no remedial group	r = NS	97.5%
" " "	(RT) - " " "	r = NS	
" " "	(R1) - plus remedial group	r = .25	66.5%
" " "	(RT) - " " "	r = .23	

NS = non-significant,
* significance at 5% level, remainder significant at 1% level

8.3.2.1 Both the correlations and the confirmed expectancy rate for the EPTB sub-sample indicated that the relationship between EPTB and tutors' ratings was much stronger for the 'no remedial' group than for the 'plus remedial' group. The correlations and contingency tables conflicted within the BCSA sub-sample. The contingency table results indicated a strong relationship with the 'no remedial' group, but the correlations indicated the reverse. The cause of this was attributed to the lack of variation in the BCSA 'no remedial' group's pre-departure proficiency. The very satisfactory level of observed improvement in English in the 'plus remedial' groups was also noted, indicating the apparent value of

remedial courses. These findings were parallel to those of the first investigation (para 6.4.1.2).

8.3.3 The sample was divided into twenty country groups, ten from the EPTB sub-sample and ten from the BCSA sample. Correlations of pre-departure EPTB Pt 1 scores or BCSA total ratings with tutors' ratings as expressed in question 1A and with their rating of total skills in question 2 were computed. Details are given in Table 8.25 in Appendix V.

Of the 40 correlations reported only 16 were significant beyond the 5% level. 11 were within the EPTB sub-sample and 5 from the BCSA sub-sample. Of the significant correlations reported only one was not greater than the whole sub-sample correlations.

No particular country or linguistic background seemed to be shared amongst those groups with the high correlations. Sudan and Egypt were both Arabic speaking countries and both groups had had positive correlations with EPTB in the first investigation. Correlations with the Turkish group were above .6, as for the first investigation, and Indonesia again produced a group with a much higher mean average age, a below average EPTB mean, a very low tutors' ratings mean, and a significant correlation. The major changes in the EPTB groups were that the Thai group was smaller and their correlation much lower, that the Iranian group achieved higher correlations and that the group from Peru, though larger, no longer obtained a high correlation. The Algerian group was again very young with a high mean EPTB score but non-significant correlations. As for the first investigation, their results were affected by their

pre-course and pre-EPTB English tuition in Britain which increased their test scores, and by the fact that most were studying for first degrees, where the demands were likely to be heaviest on English proficiency.

The relationship with BCSA reported for the Argentine group was unexpectedly strong for such a small group. The Chile group was much smaller than for the first investigation and their correlations were not significant. As in the first investigation, their assessments contrasted with those of Brazil and Jordan, where BCSA ratings had been high, at B+, but for whom tutors' ratings were below the mean. The inference was that assessors in the latter countries were interpreting the rating scales leniently, and the high levels of proficiency reported were not corroborated by their British tutors.

The pattern followed that of the first investigation (see para 6.4.2.2). Large country groups taking EPTB were most likely to exhibit strong relationships between EPTB and tutors' ratings. BCSA groups tended to be smaller and so positive relationships fewer. Evidence was now accumulating from the two investigations of the predictive power of EPTB in the first 6 countries in Table 8.25. BCSA was clearly a less predictive and less reliable instrument on an individual country basis.

8.3.4 The sample was next examined accaording to subject areas of study, each of the EPTB and BCSA sub-samples being divided into nine groups. Means, correlations and confirmed expectancies are reported in Table 8.26 in Appendix V. The principal differences

from the results obtained with the first sample were that the second investigation comparisons yielded fewer positive correlations but contained new information - the rate of confirmed expectancies for each subject group.

Thirty-four correlations were again computed (see para 6.4.3.1 above for first investigation results) - the EPTB Arts subjects group being omitted because it was so small. Only seven of the EPTB/ tutors' ratings correlations were signficant beyond the 5% level, whereas fourteen of the BCSA/tutors' ratings were. Only two zero correlations were recorded (BCSA Agriculture group), and there were no negative correlations.

The Engineering/Technology groups were larger in this investigation and again showed a strong positive relationship between pre-departure assessments and tutors' ratings. Similar satisfactory relationships were observed for both the EPTB and BCSA Social Sciences (academic) groups.

The strongest relationship revealed by EPTB was again related to the Education/TEFL students. The Arts and English Studies groups were too small for meaningful relationships to be established, but the Physical Sciences group showed only a weak, non-significant relationship. The strength of the relationships for the academic groups was, therefore, not repeated in the second investigation.

The reasons for this were not entirely clear, however. The fact that the sub-sample was one quarter smaller was thought to be one factor, and the nature of the sub-sample's improvement in English proficiency another. The whole sub-sample had reached Britain with

a slightly lower level of attainment on EPTB - 38.3 compared with 38.9 - but almost two thirds had followed pre-sessional remedial English courses. A slightly smaller proportion had been deemed by tutors to have completely inadequate English - 2.8% as compared with 3.2% for the first investigation.

By contrast satisfactory relationships were shown between BCSA and tutors' ratings for all groups, except the Agriculture and Arts groups, while the relationship for the Engineering group was weak though positive.

The rather unsatisfactory comparisons for the EPTB groups were offset, however, by contingency table expectations which had been confirmed at a rate of 80% or more for all except the Education group. The range was from 79% to 100% (English studies - a small group), with Education at 66%. The expectancy rates for the BCSA groups were lower and within a more limited range - from 73% to 86%. This continued the pattern established for the comparisons for the whole sub-samples (see para 8.2.2 above) and indicated that expectancy tables based on EPTB had been a little more accurate than those based on BCSA results.

8.3.4.1 The range of tutors' means was not so great as in the first investigation (see para 6.4.3.2 above). Neither were the discrepancies between pre-departure assessment means and tutors' ratings means for the English studies group observed to be quite as great in the second investigation. It was noted that this group, though small, came with the highest pre-departure means (EPTB - 43.4 and BCSA - 13.9/B+) and were rated above the sample mean by

their tutors. The EPTB English studies group (N=10) obtained a ratings mean of 10.3, well over half a standard deviation above the whole sample mean. The BCSA English studies group, while larger (N=35) obtained a tutors' ratings mean of 9.1 which, while just above the total sample mean, was the lowest tutors' rating mean for any BCSA sub-sample subject group, and equal to that of the social sciences (professional) group. This was considered to be further evidence that a higher standard of English proficiency is required of students studying English literature or Applied Linguistics.

It was also noted that the Agriculture group again had low pre-departure means but appeared to satisfy their tutors, while the social sciences (professional) group continued to be rated as the group with the least satisfactory proficiency. Thus two findings of the first investigation were confirmed, namely that it would appear that only a moderate standard of English is required of Agriculture students, while a higher standard is required of students in the Social Sciences (professional) courses.

8.3.5 The sub-samples were further divided according to levels of study in Britain. There was less certainty about variable of level before a student reached Britain and it may therefore have influenced the relationship between pre-departure assessments and the criterion measures in a different way. Before leaving his home country a student was normally certain of whether he was to follow a remedial English course and certain of his field of study. Sex and country of origin were fixed. But level of study was not always

clear. Frequently there would be a choice between a Master's degree or a Diploma course, a Master's by tuition or by research, and sometimes a choice between a degree course and an attachment. A decision was often taken only after a student had joined his institution. The nature of a student's English proficiency could have influenced the decision, and this is one conclusion that could be drawn from the distributions of EPTB scores in Table 8.27 below.

The major inferences to be drawn were that the highest standard of English was required for research studies, that research degree students were adjudged by their tutors to have the most acceptable level of proficiency and that the relationship between EPTB or BCSA and tutors' ratings was strongest for this group. Confirmed expectancies were also acceptably high. Standards of proficiency - particularly as assessed by EPTB - seemed to decline and relationships weaken as students took courses at lower levels, through Master's degree, diploma courses to attachments. The exception to the trend was the first degree group.

The first degree group has been included because it completes the range of levels studied even though all the subjects came from Algeria, were appreciably younger than the rest of the sample and for the most part had had extensive periods of pre-sessional English, often exceeding 3 months. Moreover, their 'pre-departure' English was often assessed during or on completion of their English courses. Nevertheless certain observations can be made. The highest standard of English as measured by EPTB was demanded of this group, with 76% of the group obtaining scores of 40.0 or above. No student was admitted with a score below the cut-off of 34.0.

Table 8.27 EPTB and BCSA distributions and comparisons with tutors' ratings by level of study

	First Degree	Master's Degree	Research (Degree)	Diploma	Academic Attachment
(a) EPTB group					
% of level with Pt 1 scores of 40 or more	76%	37%	40%	26%	18%
Pt 1/ratings (RT) r =	.19*	.35	.43	.38	.36
Confirmed expectancy rate	100%	88%	84%	83%	54%
N for level	17	88	45	115	28
(b) BCSA group					
% of level with grades of B+, A	54%	48%	67%	55%	39%
Grade/ratings (RT) r =	-.06*	.34	.58	.31	.31
Confirmed expectancy rate	69%	70%	90%	81%	74%
N for level	23	82	72	188	41

* = non-significant

This selection procedure was vindicated by the 100% expectancy rate, with no student in this group being declared completely inadequate in his English proficiency by his tutor. The group assessed by BCSA fared less satisfactorily. Only 54% were admitted with grades of B+ or A, and the expectancy rate was the next to lowest of all groups reported at 69%. Although both groups were small, the results seemed to demonstrate clearly the advisability of following the cut-off guidelines for students entering first degree courses and the superiority of the predictive power of EPTB over BCSA at this level. This superiority was maintained across the levels when expressed in terms of confirmed expectancies and even in terms of correlation coefficients.

The very low standards of English accepted at the beginning of

academic attachments was striking. Almost half the EPTB group entered with scores of under 34.0 and a quarter of the BCSA group with grades of C+ or below (see Table 8.27.1, Appendix V). It was not clear whether applicants with very poor proficiency were transferred on arrival from courses to attachments or whether standards were generally lower. In terms of tutors' ratings, however, these subjects were clearly satisfactory. Ratings at question 1 of the form followed the pattern of the whole sample with only one of the sixty-nine subjects being declared completely inadequate in English proficiency.

8.3.5.1 These findings were in line with those of the first investigation (see para 6.4.4 above). The relationships between pre-departure assessments and tutors' ratings were found to be substantial across the levels and it was concluded that the variable of level of study had a strong positive effect, particularly for those subjects pursuing research studies.

8.3.6 During the reconstitution of the sample into different experimental groups according to five different sets of variables, more than 130 comparisons were made, 100 by correlation and 34 by contingency tables. Some of the sub-groups compared were small and significant relationships could not be established. However, in the overwhelming number of cases the strong relationships already established between the pre-departure measures and tutors' ratings were confirmed. It was noted that certain variables affected the overall relationships to varying degrees. The variable of sex, when combined with subject of study, seemed to indicate a stronger

relationship for the small female sample, whose subject areas of study were predominantly academic and education. Country of origin seemed to have no appreciable effect unless country groups were large and being assessed by EPTB. The strongest effect was obtained when groups were established according to whether they had had pre-sessional remedial English courses or not. It was found that the following of such courses had an adverse effect on the relationship. The effect of subject of study was found to be minimal, but level of study did seem to be an important variable. The relationship was found to be particularly strong for those subjects following research degree courses or studies, and for those following first degree courses who had been assessed by EPTB.

8.4 Conclusions

Analysis of the results of the second investigation led to the establishing of two major sets of relationship - between the two sets of criterion measures, and between the pre-departure and the criterion measures.

It was shown that in general terms the tutors' ratings and the Communicative Proficiency Measure were yielding the same basic information on the adequacy of the students' English for their studies. An acceptable degree of concurrent validity was therefore claimed for the measures and consequently greater overall validity claimed for the English Ability Rating form completed by the tutors.

The relationship between pre-departure assessments and tutors' ratings was shown to be consistently positive and substantial as evidenced by highly significant overall correlations of .36 and by expectancy tables

yielding confirmed expectancies at a rate of over 80%, based on cross-tabulations yielding highly significant chi squares. These relationships were shown to be even stronger when pre-departure assessments were compared with performance on the Communicative Proficiency Measure, in spite of the small size of the sample. The power of EPTB Part 2 was noted and the need for further investigation of the validity of this type of test put forward.

It was possible, therefore, to give further confirmation of the hypotheses that strong and significant relationships existed between performance on EPTB or BCSA and the two criterion measures. These relationships pointed to the predictive power of both EPTB and BCSA.

It was also possible to show that this strong relationship extended to experimental groups affected by certain variables. The strength of the relationship varied, however. It was seen to be strongest for subjects following courses in Education, and to a lesser extent English and linguistic studies and Engineering, and for students studying for higher degrees - particularly research degrees. The relationship was shown to be adversely affected by attendance of remedial English classes before beginning the main course of study. This was not unexpected and presumably pointed to the general usefulness and effectiveness of such tuition.

With two investigations completed it was felt that there was enough information to reach conclusions based on both investigations and to reach conclusions on the approach adopted. These are discussed in the next and final chapter.

CHAPTER 9 CONCLUSIONS

9.0 The Context of the Study

This study has been carried out over a period of several years and it has therefore not been possible to take the most recent research and developments in language testing into account when the preliminary analysis of the situation was undertaken and when the study was being designed. It has therefore run the risk of being out of date. However, because of its duration it has been subjected to the test of perspective and time. In this final chapter, therefore, the various findings and conclusions of the different phases of the study will be drawn together and placed in the context of the early 1980s, and the general approach evaluated in the light of future likely needs and of current trends in the development of proficiency tests of English as a foreign language.

9.1 The Current Situation

There seems to be little change in the state of English language proficiency testing from that set out in Chapter 1 (paragraphs 1.4.1 to 1.4.6 above). The three perceived types of tests are still operative,

examiner-based
overall proficiency
functional proficiency based.

Although developments have taken place in the sphere of functional proficiency testing with the introduction of the English Language Testing Service by the University of Cambridge and the British Council in 1980 (ELTS, 1980), with the introduction of the Australian Second Language Proficiency Rating by the Department of Migrant Affairs, Australia, in 1980 (Ingram, 1980), and with the introduction of the new scheme of examinations in the communicative use of English by the Royal Society of Arts

(RSA, 1980), in 1981. Alongside these initiatives are the developments in the testing of French and other modern languages at lower levels of attainment in a growing number of British secondary schools inspired by the graded objectives movement (Harding, Page and Rowall, 1980).

9.1.1 Examiner-based tests of English are still administered widely, not only as teacher made achievement tests, but also in parts of established public examinations such as the Cambridge Proficiency in English, the traditional RSA examinations in English as a foreign language and the English Speaking Board examinations. These types of test may be characterised as being secretive and ad hoc. These terms may appear exaggerated but they are indicative of the fact that specifications are not always clear and may consist of only very general statements. Consequently teachers and candidates find most guidance in the study of past papers and are prepared for a marker or oral examiner who may be idiosyncratic in his setting of questions and in his reactions to whatever the candidate may say or write. There is, however, ample scope in examiner-based tests for systematic and searching setting of questions and for an analytic or controlled approach to the assessing of answers.

Overall proficiency testing tends to be coherent in its approach and format, systematic in its sampling, efficient in its use of time, and reliable. Such tests, however, often lack lack certain attributes which test consumers find desirable - relevance and attractiveness of the tasks, and a generally sympathetic approach to the learner. These are attributes that are emphasised in function-based tests.

9.1.2 There is little change in the availability of information on the validity of proficiency tests. Current debates still focus on internal validity. Oller turned from obtaining evidence of validity from external criteria related to cloze and dictation tests to seeking evidence of construct validity in the identification of a general proficiency factor in these tests (Oller, 1980). The general proficiency factor is disputed by Bachman and Palmer (1980), Vollmer (1979 and 1981) and others.

The constructors of functional proficiency tests are concerned more with attributes which are related to content and face validity, and the resources that were allocated by ETS to Pike for his study of TOEFL (Pike, 1973) were used to examine primarily aspects of internal validity - the appropriateness of test formats. The opportunity for a detailed external criterion validity study was thus lost and no further resources are known to have been allocated since. Validation of ELBA against the criterion of academic success in the University of Edinburgh has continued, and the basic trends previously reported - see para 2.5.3.4 above - have been confirmed (Howatt and Davies, 1979).

The need for more information on the external validity of proficiency tests, discussed in paras 2.6 to 2.7.3 above, remains, and the present study would not seem to have lost its relevance with the passage of time.

9.2 Evaluation of the Results of the Investigations

The approach to continuous test validation proposed in paragraphs 2.8.3 to 2.8.4 above can now be examined in order to establish the extent of

its viability in the current state of English proficiency testing.

The approach was exemplified by the establishing of five hypotheses relating to two English proficiency testing procedures. The two sets of procedures were totally different in nature. One procedure, the BCSA, was representative of the examiner-based type of language test, while the other, EPTB, was representative of the overall proficiency based type of test. Two of the hypotheses related to BCSA, two to EPTB, while the fifth related to both.

9.2.1 The first two hypotheses - paras 2.9.1 and 2.9.2 above - related to the internal consistency of the testing procedures. The tests were examined in the light of these hypotheses before the investigations involving samples of students were set up. Only when general confirmation of the first two hypotheses was obtained was it appropriate to proceed to the examination of the other three. The conclusions reached at the end of the discussion in Chapter 3 were:

(a) the construction of the English Proficiency Test Battery is consistent with the theory on which it is based, but is not basically compatible with current thinking in linguistics and language testing;

(b) the construction and administration of the British Council Subjective Assessment procedures cannot be said to be consistent with the theory on which they are based, since this is nowhere stated. The procedures are, however, consistent with their general stated aim; the general approach which can be inferred from these procedures does contain some

elements which are compatible with current thinking in linguistics and language testing.

It should be noted, however, that current thinking is not united in its conclusions on the most appropriate theory on which to base language tests. The two measures are typical of widespread current practice in language testing and much of the theory underlying EPTB is acceptable to many language testers.

9.2.1.1 The procedures adopted to reach these conclusions were straightforward. The first step was to collect all the basic documentation available on the tests. In the case of EPTB, the documentation was very complete, being mostly contained in a doctoral thesis (Davies, 1965) and including data from external criteria studies undertaken at the time of construction. Documentation on BCSA was as minimal as that on EPTB was plentiful! One basic document giving guidelines to testers was obtained and no copy of any particular examiner's test was available. It did not seem appropriate to consult any, unless a large-scale study of what a large number of testers were using had been undertaken.

The procedures for this phase of the validation were adjudged to be workable and extendable to other tests. They fulfilled two other useful functions as well. Firstly, although the internal consistency of a test may have been subject to scrutiny not only by its constructor(s) but by an expert committee as well (as in the case of EPTB), a re-examination by a different group at a later date could uncover discrepancies that were not readily apparent to those working on its construction and therefore over-

familiar with its design. Secondly, such an exercise could well prove valuable for test constructors in training.

9.2.2 The third and fourth hypotheses - paras 2.9.3 and 2.9.4 above - postulated strong positive relationships between performance by appropriate samples of students on EPTB or BCSA and appropriate criterion measures.

Two samples of overseas students in Britain were selected in two successive years. They totalled 1,812, and background data, including performance on BCSA or EPTB, was obtained for them. Two criterion measures were designed and trialled. The English Ability Rating form was sent for completion to the tutors of 1,758 students, and completed forms were returned in respect of 1,473. The Communicative Proficiency Measure was administered to 95 members of the second sample, and full data from tutors and complete test-results were obtained for 80 subjects.

Performances on the proficiency test measures and the criterion measures were compared by means of correlations and contingency tables for both samples. Highly significant positive correlations were obtained for both investigations and for both the sub-samples, EPTB and BCSA, as follows:

EPTB with tutors' ratings	.32 (first) and .36 (second)
BCSA with tutors' ratings	.35 (first) and .36 (second).

Two-way contingency tables for both investigations and for both sub-samples produced confirmed expectancy rates as follows:

EPTB with tutors' ratings	86% (first) and 82% (second)
BCSA with tutors' ratings	84% (first) and 78% (second)

These represented predictions which were from 56% to 72% more efficient than chance.

Performances on the proficiency test measures were also compared with performances on the Communicative Proficiency Measure for the second investigation. Correlations and confirmed expectancy rates were:

EPTB with CPM	.33 and 80%
BCSA with CPM	.57 and 80%

The sub-samples were small, but the correlations reported were all significant beyond the 5% level, as were the chi squares upon which the contingency tables were based. It was noted that the relationships for this small sample were just as strong as the relationships between the whole sample and tutors' ratings. These relationships were therefore considered highly satisfactory, particularly since 75% of the CPM sample had undergone remedial English tuition (Table 7.45) which had previously been found to affect the relationships adversely.

The results obtained across the two investigations were considered to amount to a demonstration of very substantial and positive relationships between both EPTB and BCSA and the criterion measures. It was further considered that these relationships could be taken as evidence of the predictive power of both EPTB and BCSA. The conclusions reached with respect to the third and fourth hypotheses and after the two investigations were:

(a) that there is a significant positive relationship between performance on EPTB by appropriate samples of students who are non-native speakers of English and performance on appropriate criterion measures;

(b) that there is a significant positive relationship between performance on BCSA by appropriate samples of students who are non-native speakers of English and performance on appropriate criterion measures.

9.2.2.1 It was noted that the English Ability Rating form had an acceptable built-in degree of reliability at .85 for both sub-samples and that the general assessment obtained by means of CPM varied substantially in only 4% (N=3) of the CPM sub-sample cases. It was therefore concluded that both the Rating form and the CPM were giving the same general information and were appropriate instruments for such a study.

It was further noted that, in view of the high rate of confirmed expectations and of the relationships between some of the pre-departure sub-tests and parts of the CPM, the latter measure could be experimented with further on other samples, ideally on a larger scale.

9.2.3 The fifth hypothesis postulated strong positive relationships between performance on EPTB or BCSA and the criterion measures for experimental groups affected by certain intervening variables. Because of the small numbers in the CPM sample it was only possible to consider relationships with one criterion measure, the English Ability Rating form. The relationships were again established by correlation and contingency tables.

The two investigation samples were divided into experimental groups according to the following variables - sex (second investigation

only), attendance at remedial English courses, country of origin, subject area of study and level of study. In all cases groups had to be further sub-divided according to the pre-departure assessment that had been taken. Thus four groups were established on the basis of sex (EPTB male, EPTB female, BCSA male, BCSA female), four for each investigation on the basis of attendance or non-attendance at remedial English courses, twenty groups according to country of origin, eighteen groups according to subject area, and ten according to level. In a number of cases groups became too small for satisfactory comparisons to be made, and such significant positive relationships as were identified were in large part identified in the larger groups.

9.2.3.1 Results were not necessarily consistent across the two investigation sample groups. However, certain trends were discerned. Firstly, the following of pre-sessional remedial English courses had an adverse effect on the relationship (paras 6.4.1.2 and 8.3.2.1 refer).

Although the correlations between EPTB or BCSA and tutors' ratings were statistically significant and positive they could not be said to point to strong relationships. Relationships were shown to be stronger with groups that had had no remedial English. The difference was more marked with EPTB groups partly because a greater proportion of the EPTB sub-sample followed remedial English courses and partly because the scores were standardised enabling more consistent judgements on attendance at such courses to be made.

Correlations with the 'no remedial' groups were significant and

positive but were lower than anticipated. It was concluded that the relationship was not particularly strong for these groups, although the trend was sustained by the EPTB group.

Subject area groups produced some strong relationships, particularly with EPTB groups in the first investigation and BCSA groups in the second (paras 6.4.3.1 and 8.3.4 refer). In general terms the relationships were found to be strongest for groups studying the more academic subjects such as English, applied linguistics, physical and biological sciences and education, including TEFL.

Further strong relationships were found with respect to levels of study, where it was noted that correlations were highest with groups following research degree courses and Master's degree courses (paras 6.4.4 and 8.3.5 refer).

This tendency was confirmed in the second investigation in respect of the female group which was small, but which had above the mean proficiency on EPTB or BCSA and whose members were studying the more academic subjects and education, including TEFL. Correlations of .39 (EPTB with tutors' ratings) and .46 (BCSA with tutors' ratings) were reported (para 8.3.1 refers).

9.2.3.2 Conclusions relating to the fifth hypothesis were not clear cut, as size of the groups seemed to be an important factor in establishing strong relationships. However, certain positive relationships appeared to have been identified and one important adverse relationship - attendance at remedial English courses. It was therefore concluded:

that there is a significant positive relationship between performance on EPTB or BCSA and performance on an appropriate criterion measure for experimental groups affected by the variables of subject area and level of study, particularly when these relate to education courses and the more academic subjects and when the courses are at research degree and other postgraduate degree levels; there is also a significant positive relationship for large single country groups who are assessed by EPTB; relationships between performance on EPTB or BCSA and an appropriate criterion measure are adversely affected when the variable of attendance of pre-sessional remedial English courses is applied.

9.3 Discussion of Issues Arising from the Results

The results of the study have led to the general confirmation of four of the hypotheses and qualified confirmation of the fifth. But they have also raised a number of issues related to proficiency testing and the validation of such testing which need to be investigated further. The question of what should be tested and what the most appropriate types of test are inevitably come to the fore. The question of the need for or the importance of this type of testing also arises, as does the question - how valid is the validation exercise?

9.3.1 Two issues are raised by this latter question - the accuracy of the findings and the validity of the criterion of 'adequacy'.

Much detailed information has been assembled and processed, and very precise statistics in the form of correlations, percentage

rates of confirmed expectancies and mean scores have been computed. Only a few variables have been taken account of in this sutdy, and many ignored. When comparing performance on English proficiency tests administered in home countries and tutors' ratings or performance on another language test, up to one year later, most intervening variables have had to be ignored. It is impossible to know what they all are, but some which must certainly be important, and to which some tutors have alluded in their comments above (para 7.4.4) are personality factors, the effect of culture change, the effect of status change, accommodation, finance and general welfare as administered by the British Council and other sponsoring bodies, British university teaching methods and living in an English-speaking environment.

The study has identified a variable with an adverse effect on the relationship - the following of pre-sessional English courses. This adverse effect is to be welcomed because presumably the courses have been beneficial in improving the subjects' English proficiency. But when this variable is added to those in the preceding paragraph it is clear that the results obtained in the study cannot claim to be precise but are indicative of <u>tendencies</u> which appear to be strong. In a validation study of this kind it would seem that the presence of strong tendencies towards certain relationships are the most that can be expected. This being so, the study has confirmed tendencies noted during the development of EPTB but not examined carefully since, and tendencies which have been hoped for in the operation of BCSA, suspected not to exist by some people, but never investigated hitherto.

9.3.2 When the criterion of adequacy was discussed earlier in this study (para 4.2.2) it was pointed out that adequacy might be defined in different ways for different courses. This has been supported by some of the findings. It was found that higher standards of proficiency in English were apparently being required of students of English and applied linguistics studies than of students of agriculture. However, the concept of the stability of the notion of adequacy seems to have been confirmed by the results of the English Rating form. Not one tutor observed that he/she was unsure of what was meant by the word adequate. Not one tutor queried his/her competence to make a worthwhile judgement. Where it was felt that no assessment could be made in respect of certain skills, this was stated and no assessment was made.

Greater reliability was also sought, and obtained, in the strength of numbers. With two large samples being investigated the influence of any erratic assessments on the general tendencies perceived was curtailed, and examination of the tutors' responses produced the very firm impression that respondents were confident of their ability to make the assessment requested and had done so in good faith and to the best of their ability. The request for further comments acted as a valuable 'safety valve' and an opportunity to make qualifications to the assessments made. It was therefore concluded that the concept of adequacy, though undefined, had proved an acceptable and meaningful criterion.

9.3.3 One of the unexpected features of the results was the close paralleling of the EPTB sub-sample results by the BCSA sub-sample

results. BCSA, as an examiner-based test, based on practice rather than on theory, and lacking in reliability, nevertheless proved as powerful a predictor, and in some cases more powerful, than EPTB. Some explanations for these need to be offered.

The power of BCSA was particularly noticeable in the correlations. Part of the explanation may lie in the method used. The total BCSA ratings scale was effectively from C to A, with intermediate + points, multiplied by four skills. When converted to numbers this became in essence a 9 point scale from 8 to 16 (eg a combination of C, B, B, B would become 11). The tutors' ratings scales at question 1A were either 3 or 4 point and at 1B six point scales. Even when notional skills ratings totals were obtained for responses to question 2, the scores tended to cluster at certain points of what was ostensibly a 13 point scale. Thus similar scales were being compared, the behaviours being assessed were the same, the four language skills, and the modes of assessment were the same - subjective impression.

This was not the case for EPTB. Obtained scores were from 34.0 to approximately 49.0. Since one decimal point was used, this was a 150 point scale with a relatively normal, though skewed, distribution. The behaviours being compared were not the same - the tests being partly of listening and reading and partly of linguistic control - and the objective method of assessment was completely different. When EPTB and BCSA both correlated with a criterion measure at eg .35, the amount of variance shared with the criterion was not great. It could have been that the variance

shared by BCSA and the criterion was influenced to a greater extent by the method of assessment adopted than was the case with EPTB.

Another possible explanation may lie in the distribution of scores of the two sub-samples. In both investigations over half the BCSA sub-sample obtained grades of B+ or A. They therefore represented the top end of the ability range and the probability of success was very high. The decision of the investigator to label subjects with grades of C+ or less as having inadequate English was arbitrarily based on the assumption that although the definitions (see Table 3.5) at C would appear to indicate a minimum adequacy, assessors were in fact 'inflating' the grades to make sure that weak or borderline cases were allowed to proceed to Britain. This assumption may well have been unjustified. In the contingency tables confirmed expectancies for the BCSA sub-sample were usually lower than for the EPTB sub-sample, and this was normally due to the number deemed (by the investigator) to be inadequate on BCSA and assessed by tutors as adequate (top right hand cell of the tables). The cut-off for BCSA was probably set too high by the investigator.

One further explanation of the strong relationships is that given at least the common assessment scales - see Table 3.5 - the majority of the testers scattered around the world were devising and administering tests to the best of their ability which enabled them to make assessments of which they felt confident. This is not an unreasonable assumption since the instructions laid down that only personnel qualified in English language teaching should administer

BCSA. Only in a minority of cases would it have been impossible to follow this instruction.

9.3.3.1 EPTB showed greater consistency over the two investigations. In terms of overall correlations both sets of measures were equally consistent, but EPTB showed consistency in the differences between the groups who had taken remedial English and those who had not. It also showed consistency in the relationships for the major country groups - Indonesia, Sudan, Turkey, Egypt, Iran and Mexico and with the more academic subject groups and levels of study groups, thereby confirming that it was performing best for the groups that it was originally intended for - those wishing to follow academic postgraduate courses.

9.3.4 One finding of the study - the very low number of subjects whose English was found by the tutors and by CPM to be inadequate - prompts the question as to the need and importance of such proficiency testing. It was noted that in the first investigation 11.2% of the EPTB sub-sample arrived in Britain with inadequate proficiency in English (Table 5.15). In the second investigation the proportion had risen to 16.9% (Table 7.11). Yet in the estimation of tutors only 3.2% were considered to have inadequate proficiency in the first investigation sample and only 2.8% in the second investigation sample (Tables 5.17 and 7.25). Since the cut-off for BCSA grades is not stated, no meaningful figures can be given, but in the first investigation 4.8% of the BCSA sub-sample were judged by tutors to have inadequate English while the figure for the second investigation was only 1.7%.

The tutors' ratings figures just quoted should be taken together with the earlier observation that the EPTB sub-sample for the second investigation arrived with generally lower proficiency than for the first investigation, while the BCSA sub-sample exhibited slightly higher proficiency in the second investigation. In addition the proportion of the whole sample following remedial English rose from 57.5% in the first investigation to 63.1% in the second, and the average period of such tuition also increased from 7.9 to 8.5 weeks.

It may be inferred from these figures that the increased remedial tuition had a more positive effect on the second investigation sample. It may also be inferred that the English proficiency of most students continued to improve once they had started their academic courses, since the apparent initial higher level of English of the second BCSA sub-sample was maintained through to the tutors' ratings. This is further supported by data reported in Tables 6.22 and 8.20. Differences in means between groups following remedial English courses and groups not following remedial courses were maintained, according to tutors' ratings, and there would seem to be little chance of students with weak proficiency catching up with those who arrived in Britain with adequate proficiency. The data is summarised below:

1st investigation		No remedial	+ Remedial	Difference
EPTB Pt 1	M	42.03	36.08	over 1 SD
Ratings (RT)	M	9.2	7.8	½ SD
BCSA total	M	15.2(A)	12.9(B)	over 1 SD
Ratings (RT)	M	10.0	8.4	2/3 SD
2nd investigation				
EPTB Pt 1	M	41.5	36.6	1 SD
Ratings (RT)	M	10.1	8.8	¾ SD
BCSA total	M	15.1(A)	12.3(B)	over 1 SD
Ratings (RT)	M	10.5	9.1	¾ SD

It is probable that the gap would widen if no remedial tuition were given to the weaker students. Further experimentation with varying lengths of remedial tuition should indicate if the periods of such tuition should be reduced or maintained.

The question of the importance of English proficiency testing in the context of study in Britain has still not been answered, however. Such testing seems nevertheless to be essential, since course organisers need to have some indication as to whether an applicant - at the time of application - possesses English proficiency which is judged to be almost adequate but susceptible to satisfactory improvement, or proficiency which is judged to be totally inadequate. Carefully worked out tests such as EPTB and the majority of BCSA procedures appear to be necessary hurdles which have to be surmounted by all non-native speaking applicants. What does seem to have to be resolved, however, is the lowest acceptable cut-off score on a proficiency test which, after the variables of remedial English, subject and level of study have been considered, can predict with sufficient probability that a minimal adequacy in English for following a course can be achieved. No one cut off is likely to provide the solution, but the range for EPTB, at least, would seem to lie somewhere between 32 and 36, depending on the type and level of course and the remedial English arranged. The BCSA assessment scales are not sufficiently sensitive, but it would seem that applicants with grades of level C could be accepted.

9.3.5 Another issue raised by the present study is the need for tests of English for special purposes. In the light of the findings

related to intervening variables it could be suggested that students intending to follow agriculture or medicine or professional courses in the general area of social sciences might be better served with different tests. Yet although the relationships for groups in certain fields of study were strong and in others not significant, the level of course seemed to be the more important variable. And level of course is differentiated by the kinds of study skills required and not by the subject matter. A case could be put therefore for differences in testing procedures which may be geared more to study styles and skills. There is no strong evidence in this study that the tests examined were inappropriate for any particular group. There was evidence, however, that subjects in certain groups, notably those proceeding on academic attachments, were being accepted with very low levels of English proficiency, which were mostly acceptable to the tutors. Tutors' reactions would seem to indicate the need for a robust and stable test for which information could be built up systematically over a period of years.

9.3.6 A more important question would appear to be related to the desirability of testing communication. Munby (1978) has highlighted the need for designing communicative courses. One of the aims of ELTS is to find out if a student's proficiency is adequate for his communicative needs. In a recent book Carroll (1980) stressed the importance of testing communication. However, the notion of communication is as ill-defined as the notion of language proficiency, although certain principles would seem to apply. Information and

ideas have to be communicated by a subject in a variety of situations in the spoken mode (monologue, dialogue, discussion) and in the written mode. The subject must also be observed to be capable of understanding information and ideas conveyed in those modes.

Recent communicative tests in English have made use of procedures that require subjects to speak with other subjects, interviewers or assessors. They also require subjects to write on specific topics with varying degrees of control and in different formats. They also require subjects to process information perceived in written or spoken texts. EPTB does not apparently possess any communicative sub-tests, but with its interview for assessing speaking and its writing component BCSA possesses elements which test communication in some of the terms outlined above.

A look at the analysis of results for the second investigation is appropriate. The sub-tests which show the strongest positive relationships with the criterion measures are the writing test assessments and the speaking assessments. Writing (D4) correlated with overall tutors' ratings at .36, which was the BCSA total/ ratings total correlation (Table 8.6.1 in Appendix V). In the much smaller CPM sub-sample writing correlated with CPM total at .535 and with writing and speaking combined at .578 (Table 8.13, Appendix V). The speaking sub-test (D2) correlated at .543 with total CPM and at .679 with combined writing and speaking. The figures for total BCSA and CPM were only marginally stronger at .568 and .662. For the EPTB sub-sample the relationship between performance on oral (DO) or writing (DW) and total tutors' ratings (RT) was just as strong as the EPTB Pt 1 relationship with tutors' ratings. Correlations

obtained were DO/RT r = .36, DW/RT r = .35 and EPTB Pt 1/RT r = .36 (Table 8.5.1, Appendix V).

These results suggest that speaking and writing tests are potentially powerful predictors of adequacy in English.

9.3.6.1 Before reaching a final conclusion the performance of one more sub-tests should be considered - the speed reading test in EPTB Pt 2. It produced the strongest relationship of any sub-test or total test with the criterion measures. Although less than 40% of the EPTB sub-sample took the test (D5), it correlated at .4 with total tutors' ratings, more than .04 higher than EPTB total, writing and oral (Table 8.5.1, Appendix V). Only 13 members of the CPM sample had taken the test but the relationships with CPM and tutors' ratings were even stronger for that group:

D5 with CPM overall	r =	.69
D5 with writing/speaking	r =	.86
D5 with writing only	r =	.85 (Table 8.12)
D5 with tutors' ratings (RT)	r =	.7 (Table 8.14)

One conclusion from these results is that further investigations of the predictive power of this test should be carried out using larger samples.

The reason for this power is not easy to determine. However, it is the EPTB sub-test which seems to make most demands on a subject's language control, which in this case appears to draw on a wide range of language competencies, both discrete and combined. The intrusive word test is like a cloze test in reverse drawing on similar skills and presumably taxing a subject's overall proficiency. The relationship between performance on the speed reading and on the

CPM cloze tests was also very strong, r = .757 (Table 8.12, Appendix V).

It is not clear whether this test can be called a communicative test. It has correlated very highly with cloze, but it may have a stronger relationship with productive skill tests than cloze does. (The CPM cloze only correlated with CPM writing and speaking at .5.

9.3.6.2 The foregoing discussion points to the desirability of including communicative test tasks in a proficiency test battery. Tests of oral and written communication ability are clearly important while, if the scope of communicative testing is extended to include advanced tests of language control, there could be intrusive reading tests, or cloze tests, or both. In this way communicative proficiency and overall proficiency skills would be tested.

9.3.7 Is there an appropriate type of proficiency test? The study did not set out to find an answer to this question but some observations can be made as a result of the study, although the answer will depend on the theory of proficiency that is adopted. Sub-tests of linguistic control, language control and of communication have been examined. These relate to what have been referred to above as overall proficiency based and functional proficiency based tests. A proficiency test of English as a foreign language on whose results important decisions are taken about individuals' future activities should test elements of both types

of proficiency. The overall proficiency test showing the strongest relationships with criterion measures was the speed reading test (EPTB Pt 2). Test 4, the grammar sub-test in EPTB Pt 1, showed satisfactory relationships with tutors' ratings - r = .33 (D4 with RT in Table 8.5.1) - and with CPM results as follows:

> with cloze r = .38, with writing/speaking r = .33 and with overall CPM r = .42. (Row D4 in Table 8.12, Appendix V.)

A case can be argued for a multiple choice test of suitably sampled grammatical items in the Battery. Functional proficiency can be tested through speaking and writing tests as discussed in paras 9.3.6.1 and 9.3.6.2 above.

No satisfactory relationships were established with the listening sub-tests investigated in the study. However, further investigations with more recent types of language control listening tests might point to a suitable listening test to add to the reading, grammar, speaking and writing tests already identified. These tests could form the nucleus of an appropriate English proficiency test battery.

9.4 Evaluation of the Approach

The study was carried out in four stages as planned, see para 2.10 above:

(a) assessment of the internal consistency of the tests

(b) identification of an appropriate sample of students

(c) designing the criterion measures, and

(d) analysing the data.

Stages (b) and (d) were repeated twice in two investigations, while stage (c) was extended over the period of the first investigation.

The approach proved workable, given the cooperation of university and other tutors and of the students themselves. Assuming that this cooperation will be forthcoming in the future the approach could be applied in a follow-up validation study of any other proficiency test. It has proved workable with an examiner-based set of procedures and with an overall proficiency test. It should prove equally workable with a functional based proficiency test, since this type of test contains elements similar to elements of both types of assessment investigated in this study.

It has been pointed out that only tendencies can be identified in view of the many intervening variables affecting relationships between pre-departure assessments and criterion 'in course' measures. The criterion measures developed have shown signs of sufficient robustness that they cculd be used in future studies, and the methods of analysis - correlation and contingency tables - have proved adequate in providing two methods of comparison during analysis of the data that can indicate very definitely where the tendencies lie.

Future application of the approach need not be so time-consuming as the present study has been, since the criterion instruments are designed and available. Any modifications that might be deemed necessary because of characteristics of the tests being investigated would have to be monitored, however.

When applied to a new proficiency test the approach could now be carried out in four stages as follows:

(a) assessment of the internal consistency of the test
(b) identification of an appropriate sample of students
(c) application of the criterion measures
(d) analysis of the results

There would be a time-lag between stages (b) and (c), and application of the Communicative Proficiency Measure could be demanding on staff time. However, the procedure would be suitable for adoption by large institutions such as universities for a continuous validation of the proficiency tests used by their non-native English speaking entrants. Data could be stored and analysed by means of simple computer programmes, eg Statistical Package for the Social Sciences (McGraw Hill), and clerical staff time required would not be great. Specialist analysis would be required only periodically and then for only brief periods of time. A system could be established in which involvement in such a validation process could be programmed into the annual workloads of admissions and departmental staff.

The general conclusion reached on the approach proposed and trialled in this study is that it has proved a valid and practical approach which can be extended to the validation studies of other similar proficiency tests. It is also an approach which could be followed and whose principles could be employed in the external validation of language proficiency measures designed to assess proficiency in English or foreign languages for purposes other than following training and academic courses, eg for use in specific occupations.

REFERENCES

AACRAO/ALD (1971) Participant Selection and Placement Study. Washington DC, American Association of Collegiate Registrars and Admissions Officers.

Alderson, J C (1978) A Study of the Cloze Procedure with Native and Non-native Speakers of English. University of Edinburgh: PhD Thesis.

Allen, J P B and Corder, S P (eds) (1974) Techniques in Applied Linguistics. The Edinburgh Course in Applied Linguistics, Volume 3. London: Oxford University Press.

Anderson, J (1971) "A technique for measuring reading comprehension and readability" English Language Teaching 25 pp 178-182.

Briere, E J (1971a) "Are we really measuring proficiency with our foreign language tests?" Foreign Language Annals 4 4 pp 385-391.

1971(b) "Discrete versus non-discrete testing in English language proficiency" in Cherrier (ed) pp 133-139.

British Council (1972) Overseas Standing Instruction "Assessments of Proficiency in English". London, restricted to internal British Council use.

(1973) EPTB Form C Scoring Instructions - "Comparative Statistics". London, restricted to internal British Council use.

(1975) Statistics of Overseas Students in Britain 1973/74. London: British Council.

(1980) Statistics of Overseas Students in the United Kingdom 1978/79. London: British Council.

Buros, O K (ed) (1965) Sixth Mental Measurements Yearbook Volume 1. New York: Gryphon Press.

(1972) Seventh Mental Measurements Yearbook Volume 1. New York: Gryphon Press.

Byers, P P (1969) Predicting TOEFL scores on the basis of English grades on the Hong Kong School Certificate and Certificate of Education. Hong Kong: Institute of International Education, unpublished manuscript.

Carroll, B J (1978) An English Language Testing Service: Specifications. London: British Council.

(1980) Testing Communicative Performance. Oxford: Pergamon.

Carroll, J B (1961) "Fundamental considerations in testing for English language proficiency of foreign students" in Testing the English Proficiency of Foreign Students. Washington DC: Center for Applied Linguistics and Modern Languages Association of America pp 30-41.

(1962) "Prediction of success in intensive foreign language training" in Training, Research and Education. University of Pittsburgh pp 95-116.

(1968) "The Psychology of language testing" in Davies (ed) pp.46-49.

(1973) "Foreign language testing: will the persistent problems persist?" in Testing in Second Language Teaching: New Dimensions. Dublin: ATESOL pp 6-17.

Carroll, J B and Sapon, S M (1959) The Modern Language Aptitude Test. New York: The Psychological Corporation.

Cazden, C B, John, V P and Hymes, D (1972) Functions of Language in the Classroom. New York: Teachers College Press.

Chalmers, P M (1964) Interpretation of the Test of English as a Foreign Language (TOEFL) scores. Boston: Massachusetts Institute of Technology, unpublished manuscript.

Chaplen E F (1970) The Identification of Non-native Speakers of English Likely to Under Achieve in University Courses through Inadequate Command of the Language. University of Manchester: PhD thesis.

Chase, C I (1972) Review of the Test of English as a Foreign Language (TOEFL) in Buros (ed) pp 550-552.

Cherrier, R (ed) (1971) English Language Testing: Report of the RELC Fifth Regional Seminar. Singapore: RELC.

Clark, J L D (1972) Foreign Language Testing: Theory and Practice. Philadelphia, Pa: Center for Curriculum Development.

(1975) "Theoretical and technical considerations in oral proficiency testing" in Jones and Spolsky (eds) pp10-24.

(1980) Language testing research at Educational Testing Services. Paper presented at the second IUS international symposium on language testing in Darmstadt.

Cooper, R L (1972) "Testing" in Allen, H B and Campbell, R N (eds) Teaching English as a Second Language: A Book of Readings 2nd ed. New York: McGraw-Hill.

Cowie, A P and Heaton, J B (eds) (1977) English for Academic Purposes: Papers on the Language Problems of Overseas Students in the UK. University of Reading: British Association of Applied Linguistics.

Centre de Recherches et d'Etudes pour la Diffusion du Français (CREDIF) (1967) Voix et Images de France. Paris: Didier.

Cronbach, L J (1960) Essentials of Psychological Testing 2nd ed. New York: Harper and Row.

Darnell, D K (1970) "Clozentropy: a procedure for testing English language proficiency of foreign students" Speech Monographs 37 1 pp 36-46.

Domino, G (1966) Academic achievement and English proficiency of foreign students as predicted by the Test of English as a Foreign Language. Fordham University: unpublished.

Davies, A (1964) English Proficiency Test Battery Short Version, Form A. London: British Council.

(1965) Proficiency in English as a Second Language. University of Birmingham: PhD thesis.

(1965a) English Proficiency Test Battery Short Version, Form B. London: British Council.

(1967) "The English Proficiency of Overseas Students" British Journal of Educational Psychology 37 2 pp 165-174.

(ed) (1968) Language Testing Symposium. London: Oxford University Press.

(1973) "Language proficiency testing and the syllabus" in Testing in Second Language Teaching: New Dimensions. Dublin: ATESOL pp 18-26.

(1978) "Survey article: language testing" in Language Teaching and Linguistics: Abstracts 11 3 and 4 pp 145-160 and 215-232.

Davies, A and Alderson, J C (1977) English Proficiency Test Battery Short Version, Form D. London: British Council.

Davies, A and Moller, A (1973) English Proficiency Test Battery Short Version, Form C. London: British Council.

van Ek, J A (1975) The Threshold Level. Strasbourg: Council of Europe.

English Language Teaching Development Unit (ELTDU) (1975) English Language - Stages of Attainment Scale. Oxford University Press, prepared for Aktiebolaget Svenska Kullagerfabriken.

English Language Testing Service (ELTS) (1980) User Handbook. University of Cambridge Local Examinations Syndicate and the British Council.

Fountain, R (1974) A case for the dictation test in selection of foreign students for English medium study in New Zealand. University of Edinburgh: Department of Linguistics.

Gradman, H L and Gaies, S J (1975) Reduced redundancy and error analysis: a study of selected performances on the "noise test". Paper presented at the 4th AILA congress in Stuttgart.

Gradman, H L and Spolsky, B (1975) "Reduced redundancy testing: a progress report" in Jones and Spolsky (eds) pp 59-66.

Groot, P J M and Harrison, A (1978) A Specimen Test for Threshold English: Manual. Strasbourg: Council of Europe.

Gue, L R and Holdaway, E A (1973) "English proficiency tests as predictors of success in graduate studies in education'. Language Learning **23** 1 pp 89-104.

Guilford, J P and Fruchter, B (1973) Fundamental Statistics in Psychology and Education. Tokyo: McGraw-Hill.

Halliday, M A K (1973) Explorations in the Functions of Language. London: Edward Arnold.

Harding, A, Page, B and Rowall, S (1980) Graded Objectives in Modern Languages. London: Centre for Information on Language Teaching and Research.

Harris, D P (1967) English Testing Guidebook. Washington DC: American Language Institute, Georgetown University.

(1969) Testing English as a Second Language. New York: McGraw-Hill.

(1972) A brief consideration of the findings of the AACRAO/ AID study concerning English language training and testing. Georgetown University: English Language Institute.

Harris, D P and Palmer, L A (1970) Comprehensive English Language Test for Speakers of English as a Foreign Language. New York: McGraw-Hill.

Heaton, J B (1975) Writing English Language Tests. London: Longman.

Heaton, J B and Pugh, A K (1974) A study of the relationship between scores obtained by overseas students on a Test of English Proficiency and the examination results in their university courses. University of Leeds: School of Education.

Hindmarsh, R X (1977) "An overview of English language testing overseas" in Cowie and Heaton (eds) pp21-24.

Horst, P (1966) Psychological Measurement and Prediction. Belmont, Ca: Wadsworth.

Howatt, A P R and Davies, A (1979) ELBA testing 1973-77: Interim Report on analysis of results. University of Edinburgh: Department of Linguistics.

Howell, T M (1975) Survey of British based examinations for overseas students. University of London: Birkbeck College, MA report.

Hymes, D (1972) Introduction to Cazden, John and Hymes pp xi-lvii.

Ibé, M D (1974) Change in the English language proficiency of RELC four-month course participants. Singapore: Regional English Language Centre, a report.

Ingram, D E (1980) "The Australian Second Language Proficiency Ratings" Paper presented at the Regional Seminar on Evaluation and Measurement of Language Competence and Performance in Singapore.

Ingram, E (1964) English Language Battery (ELBA). University of Edinburgh: Department of Linguistics.

(1970) Manual for the English Language Battery. University of Edinburgh: Department of Linguistics.

(1971) "The marking of English compositions". University of Edinburgh: Department of Linguistics.

(1973) "English standards of foreign students". University of Edinburgh Bulletin 9 12 pp 4-5.

(1974) "Language Testing" in Allen and Corder (eds) pp313-343.

Jakobovits, L A (1970) Foreign Language Learning: a Psycholinguistic Analysis of the Issues. Rowley, Mass: Newbury House.

Jones, R L (1975) "Achieving objectivity in subjective language tests". Paper presented at the 4th AILA congress in Stuttgart.

Jones, R L and Spolsky, B (eds) (1975) Testing Language Proficiency. Arlington, Va: Center for Applied Linguistics.

Jordan, R R (1977) "Identification of problems and needs: a student profile" in Cowie and Heaton (eds) pp 12-20.

Krowicz, L and Garcia-Zamor, M (1975) Tables appended to a paper on a functional language testing programme at the World Bank presented at the 4th AILA congress in Stuttgart.

Lado, R (1957) Linguistics Across Cultures: Applied Linguistics for Language Teachers. Ann Arbor: University of Michigan Press.

(1961) Language Testing: the Construction and Use of Foreign Language Tests. London: Longman.

Lyons, J (1968) Introduction to Theoretical Linguistics. Cambridge: Cambridge University Press.

Maxwell, A (1965) A comparison of two English as a foreign language tests. University of California (Davis).

Mialaret, G and Malandain, C (1962) Test CGM 62. Paris: CREDIF and Didier.

Moller, A (1975) "Testing the robustness of a proficiency test." Paper presented at the 4th AILA congress in Stuttgart.

(1977) "A case for a crude test overseas" in Cowie and Heaton (eds) pp 25-33.

Morris, B (1967) International Community?. London: National Union of Students.

Morrow, K (1977) Techniques of Evaluation for a Notional Syllabus. London: Royal Society of Arts.

Munby, J (1978) Communicative Syllabus Design. Cambridge: Cambridge University Press.

Oller, J W (1971) "Dictation as a device for testing foreign language proficiency" English Language Teaching 25 3 pp 254-9.

(1973) "Cloze tests of second language proficiency and what they measure" Language Learning 23 1 pp 105-118.

(1979) Language Tests at School. London: Longman.

(1980) "Pragmatics and language testing". Paper presented at the Regional Seminar on Evaluation and Measurement of Language Competence and Performance in Singapore.

Oller, J W, Bowen, J D, Ton That Dien and Mason, V W (1972) "Cloze tests in English, Thai and Vietnamese: native and non-native performance". Language Learning 22 1 pp1-15.

Oller, J W, Irvine, P and Parvin, Atai (1974) "Cloze, dictation and the test of English as a Foreign Language" Language Learning 24 2 pp245-252.

Oller, J W and Streiff, V (1975) "Dictation: a test of grammar based expectancies" in Jones and Spolsky (eds) pp 71-82.

Oller, J W and Perkins, K (1980) Research in Language Testing. Rowley, Mass: Newbury House.

Page, B (1978) Graded Examinations Modern Languages 59 2 pp 97-101.

Palmer, A and Bachman, L (1980) "Basic concerns in test validation". Paper presented at the Regional Seminar on Evaluation and Measurement of Language Competence and Performance in Singapore.

Perren, G E (1970) "Examinations in English as a Foreign Language" CILT Reports and Papers 4.

Petersen, C R and Cartier, F A (1975) "Some theoretical problems and practical solutions in proficiency test validity" in Jones and Spolsky (eds) pp 105-114.

Pike, L (1973) An evaluation of present and alternative item formats for use in the Test of English as a Foreign Language: draft report. Princeton, NJ: Educational Testing Service.

Pilliner, A E G (1974) A visit to CIEFL, Hyderabad. Report to the British Council. University of Edinburgh: Godfrey Thomson Unit for Academic Assessment.

Pilliner, S (1965) A comparison of two English language proficiency test batteries. University of Edinburgh: Department of Linguistics.

Pitcher, B and Ra, J B The relation between scores on the Test of English as a Foreign Language and ratings of actual theme writing. Statistical Report 67-9. Princeton, NJ: Educational Testing Service.

Royal Society of Arts (RSA) (1980) Examinations in the Communicative Use of English as a Foreign Language: Specifications and Specimen Papers. London.

Seliger, H W (1975) "Two experiments in foreign language testing and acquisition". Paper presented at the 4th AILA congress in Stuttgart.

Sen, A (1970) Problems of Overseas Students and Nurses. Slough: National Foundation for Educational Research.

Sharon, A (1971) Test of English as a Foreign Language as a moderator of Graduate Record Examinations score in the prediction of foreign students' grades in graduate schools. Princeton, NJ: Educational Testing Service.

Smith, A de W (1973) Generic Skills for Occupational Training. Prince Albert, Saskatchewan: Department of Manpower and Immigration.

Spolsky, B (1967) "Do they know enough English? Some notes on the problems of assessing the proficiency in English of foreign students" in Wigglesworth (ed) pp 30-43.

(1968) "Language testing - the problem of validation" in TESOL Quarterly 2 2 pp 88-94.

(1975) "Language testing: art or science?" Main lecture delivered at the 4th AILA congress in Stuttgart.

Spolsky, B, Bengt Sigurd, Masahito Sako, Edward Walker and Catherine Arterburn (1968) "Reduced redundancy as a language testing tool" in Language Learning 18 special issue.

Stevenson, D K (1975) "Construct validation and language proficiency measurement". Paper presented at the 4th AILA congress in Stuttgart.

Taylor, C (1975) Language Testing: The Dictogloss Pre-test. Glebe, New South Wales: Australasian Medical Publishing Co.

Taylor, W L (1953) "Cloze procedure: a new tool for measuring readability". Journalism Quarterly 30 pp 33-42.

Test of English as a Foreign Language (TOEFL) (1970) Interpretive Information. Princeton NJ: Educational Testing Service.

(1973) Manual for TOEFL score recipients. Princeton, NJ: Educational Testing Service.

(1981) Test and Score Manual. Princeton NJ: Educational Testing Service.

Thorndike, R L and Hagen, E P (1969) Measurement and Evaluation in Psychology and Education 3rd ed New York: Wiley.

Trim, J L M (1977) Report on some Possible Lines of Development of an Overall Structure for a European Unit/Credit Scheme for Foreign Language Learning by Adults. Strasbourg: Council of Europe.

University of Cambridge Local Examinations Syndicate (UCLES) (1973) Cambridge Examinations in English, Changes of Syllabus 1975. Cambridge.

(1980) Cambridge Examinations in English Survey. Cambridge.

Valette, R M (1967) Modern Language Testing - a Handbook. New York: Harcourt, Brace and World.

(1977) Modern Language Testing 2nd ed. New York: Harcourt Brace Jovanovich.

Vaughan James, C and Rouve, S (1973) Survey of Curricula and Performance in Modern Languages 1971-2. London: Centre for Information on Language Testing and Research.

Vollmer, H J (1979) "Why are we interested in General Language Proficiency?". Paper presented at the 1979 German International Symposium on language testing in Hürth.

(1981) "Receptive vs productive competence? - Models, findings and psycholinguistic considerations in L2 testing". Paper presented at the 6th AILA congress in Lund.

Whiteson, V (1972) "The correlation of auditory comprehension with general language proficiency". Audio-Visual Language Journal 10 2 pp 89-91.

Wigglesworth, D C (ed) (1967) Selected Conference Papers of the Association of Teachers of English as a Second Language. Los Altos, Ca: National Association for Foreign Student Affairs.

Wilds, C P (1975) "The oral interview test" in Jones and Spolsky (eds) pp 29-38.

A STUDY

IN THE VALIDATION

OF PROFICIENCY TESTS OF

ENGLISH AS A FOREIGN LANGUAGE

VOLUME II

APPENDICES

APPENDIX I CRITERION MEASURES - PRELIMINARY VERSIONS

Section 1 English Ability Rating

- Trial version of English Ability Rating form
- Revised version of English Ability Rating form
- Key for completing English Ability Rating form
- Letter to tutors

TRIAL VERSION

Name ..... Field of study .....

ENGLISH ABILITY RATING

1. Please answer the following question by putting X in the appropriate box in column A and another X in the box by the most appropriate statement in column B.

Is the general ability in English of this student adequate for undertaking specialised studies or research in his field of study?

A		B	
☐	More than adequate	☐	Shows native speaker ability.
		☐	Clearly a non-native speaker because of minor faults in English usage. But this does not handicap him in his studies.
☐	Adequate	☐	Makes many mistakes in English, but this constitutes only a minor handicap for him in his studies.
		☐	Shows many weaknesses in English usage but his English ability can be considered just adequate for his studies. A higher standard is desirable.
☐	Not adequate	☐	Shows considerable deficiencies in English usage, which constitutes a handicap for him in his studies. A higher standard is necessary.
		☐	Shows very little ability in English and is well below a satisfactory standard.

2. Please consider this student's ability in the following language skills and where possible give your rating of his ability in each skill by putting X at an appropriate point along the scale from zero ability to native speaker ability.

	Not adequate		adequate		More than adequate
Ability to understand spoken English	: .	:	.	:	. :
Ability to speak English	: .	:	.	:	. :
Ability to understand written English	: .	:	.	:	. :
Ability to write English	: .	:	.	:	. :

3. Do you think this student has shown any improvement in his English ability since October 1973?

☐ considerable improvement ☐ a little improvement ☐ no improvement

Date ..... Tutor .....

Student's Name ..... Field of Study .....

Country ..... University in U.K. .....

ENGLISH ABILITY RATING

1. General ability in English
(Put X in the appropriate box in column A and in column B)

A	B
☐ completely adequate	☐ Shows native speaker ability
	☐ Clearly a non-native speaker because of minor faults in English usage, but this does not handicap him/her in his/her studies
☐ just adequate	☐ Makes many mistakes in English, but this constitutes only a minor handicap for him/her in his/her studies
	☐ Shows many weaknesses in English usage but his/her English ability can be considered just adequate for his/her studies. A higher standard is desirable
☐ not adequate	☐ Shows considerable deficiencies in English usage, which constitutes a handicap for him/her in his/her studies. A higher standard is necessary
	☐ Shows very little ability in English and is well below a satisfactory standard

2. Individual language skills
(Put X at an appropriate point along the scale for each skill)

	completely adequate — adequate — inadequate
Ability to understand spoken English	
Ability to speak English	
Ability to understand written English	
Ability to write English	

3. Improvement in English ability since October 1973

☐ considerable improvement ☐ a little improvement ☐ no improvement

4. Tuition in English since October 1973

☐ Yes ☐ No ☐ Don't know

5. Qualification aimed at, on attachment, or research only

.....

6. Please use the back of this form for further comments

Date ..... Tutor .....

ENGLISH ABILITY RATING

Key for completing the accompanying English ability rating form

- Please complete one form per student
- The information requested below should be supplied in the boxes or spaces on the accompanying rating form

1. Is the general ability in English of the student (i.e. present ability, not potential) adequate for undertaking specialised studies or research in his/her field of study?

   Please answer this question by putting X first in the appropriate box in column A and then in the appropriate box in column B.

2. Please consider the student's ability in the following language skills -
   ability to understand spoken English
   ability to speak English
   ability to understand written English
   ability to write English
   and where possible give your rating of his/her ability in each skill by putting X at an appropriate point along the scale from completely adequate (i.e. native speaker ability) to inadequate (i.e. little or no ability).

3. Do you think the student has shown any improvement in his English ability since October 1973?

4. Has the student received any tuition in English since October 1973?

5. Please indicate which degree, diploma or other qualification the student is working for, or whether he/she is on attachment or doing research only.

6. Please use the reverse of the rating form for any further comments you may wish to make.

- Please check that the student's field of study is correctly given at the head of the rating form and put the date and your own signature at the foot of the form.

Department of Linguistics
University of Edinburgh
Adam Ferguson Building
George Square
Edinburgh EH8 9LL

March 1974

English Language Ability of Overseas Students

I should be grateful for your assistance in a research project which I am undertaking under the supervision of Dr. Alan Davies in the Department of Linguistics at the University of Edinburgh. With this letter you will find some English Ability Rating forms in respect of overseas students who are studying under your supervision or in your department. There is also a key explaining how to complete the forms. It would be of immense value to me if you, or the member of the departmental staff who has most contact with the students listed, could complete the forms and return them to me at the above address by 31st March 1974, or as soon as possible thereafter.

The purpose of the research is to obtain assessments of the English ability of those overseas students who are administered by the British Council, who have been in Britain for some months, and whose English was assessed before they left their home country. It is hoped that the results of this research will lead to

1. more efficient procedures in the selection of overseas students for awards to universities and colleges in Britain
2. improvement in the assessment of candidates' English
3. more specific indications as to the type of remedial English work that some students will need.

I am sure you will agree that this is an important area of investigation and hope that you will be willing to spend a few minutes of your time to assist me.

I am a member of the British Council staff on detachment to the University of Edinburgh to undertake further research into the testing of English proficiency of students coming to Britain to study. This work is supported by the English Teaching Division of the Council as well as the Technical Assistance Training and Scholarships departments who have made the background material available from which I have selected the present sample of students.

The rating form has been designed for completion by supervisors who are not specialists in English language teaching. However, there may be an item, e.g. in paragraph 2, that you may feel unable to complete. If that is the case, please complete as much of each form as is possible. There is space for further comments on the reverse. It goes without saying that all the assessments given will remain confidential. Assessments of individuals will not be divulged, nor will the identities of assessors. If you require any further information, please write to me.

I look forward to receiving the completed forms and thank you in advance for your co-operation.

Yours sincerely,

A. D. MOLLER.

Section 2 Communicative Proficiency Measure - trial version

- Dictation

- Essay

- Reading Test Passages

- Reading Test Passages - original texts

DICTATION

(Note: Numbers in the gaps denote the length of pause in seconds.)

In a minute I shall dictate these developments to you. You should write during the pauses. I shall not repeat any sentences or part of a sentence. No punctuation will be given. You should supply the punctuation as best you can. You may write any numbers in figures. The proposed developments are numbered from one to four. Now, please write as I dictate to you.

1. We now know 4 which types of project are successful, 10 and we can therefore provide 10 an outline programme at the beginning of the year 16 which can be discussed 8 by teachers and students together. 12

2. Arrangements for borrowing university material 15 must be more systematic: 9 collections of items 4 specifically for primary school use 10 will be set aside in the zoology department. 14

3. It is clear 2 that teachers would like more visual aids 12 in the form of charts, 6 posters, models, slides etc. 8 A small collection of these 8 could be built up if funds were available. 20

4. If more schools join our scheme, 8 we may need to extend our student numbers. 12 It is not yet clear 4 what the maximum number required may be. 13 Our main source of students 8 will remain the first year medical and biological classes. 16 At the moment it is difficult for students to work in schools 16 beyond their second year. 8

ESSAY 30 minutes

Write four or five paragraphs about a part of your work this year that has interested you very much. Describe this work and explain why you have found it interesting and enjoyable.

The following are notes and suggestions for your guidance.

- The aspect of your work that you write about may be

  a course of lectures or readings on a particular topic
  a project that you have participated in
  a piece of research that you have carried out
  an experiment that you have conducted
  some teaching material that you have compiled
  a visit or series of visits

- In your description you should include

  the purpose of this particular piece of work
  how the work was carried out or presented
  what it involved for you - what you had to do
  what the result or conclusion was

- The explanation of your interest in this work may be influenced by the fact that it was

  something new to you
  something you had always wanted to do
  something that will be very useful to you in your country
  another way of approaching a familiar task or problem

You will have to be brief as you have only 30 minutes. The accuracy of your facts is not as important as your fluency and style of expression. You will not lose credit for not finishing within the time, but you are asked to leave enough time to give some reasons for your interest in the work described.

This passage is about a comet that has recently been discovered. The comet is noteworthy for its brightness and its comparative closeness to the sun. There is a lot of speculation as to whether we shall be able to see it in daylight later this month.

The most recent predictions strongly discount the possibility of seeing the ..... in full daylight; they put ..... maximum brightness at close to ..... of Venus (which by the ..... is a brilliant object at ....., visible early in the evening ..... it sets in the south-west). ..... our northern latitude does not ..... the viewing of this particular ....., which rises and sets at ..... small angle to the horizon ....., appears ..... low in the sky.

The ..... was photographed at the Observatory ..... its brightness was lower than ..... of the faintest naked eye ..... . It will,however, continue to ..... brighter as it approaches the ..... , and by the middle of ..... month it will be readily ..... to the naked eye in ..... morning sky, a few degrees ..... the horizon in a direction ..... of south about half an ..... before sunrise. When at its ..... luminous, on the 28th, it ..... be only one degree away ..... the sun, and unless it ..... a brilliant daylight comet after ..... , it will be invisible to ..... .

Start .....
Finish.....

AC 30

This passage was written by a professor of anthropology in an introduction to a book, written by one of his former students, abqut the customs of an African tribe.

The chapter on magic is especially important and valuable because of the abundance of texts, of the details given as to the ritual and ingredients employed, and the inside information on the psychology of act and situation alike. Some anthropologists ..... question here the reinterpretation of ..... real processes which underlie magic. "..... can safely be said that ..... (i.e. magic) is one way of ..... thoughts telepathetically from one mind ..... another," we are told. "It ..... through concentration, the magician ..... the possessor of love magic ..... able to penetrate into the ..... mechanism of the person with ..... he desires to establish communication. ..... this form the magician's suggestions ..... easily transmitted by means of ..... to the brain, and then ..... the mind.

I submit that ..... author would still have to ..... some evidence as to how ..... 'vibrations' are produced, how they ..... on the brain, and then ..... the mind. But the author ..... all criticism when he tells ..... "There is something in it ..... can be classified as occultism, ..., as such, it cannot be ..... as merely superstition."

Start .....
Finish.....

BC 25

When he first went to Cambridge as a young man, Chapman was shown round the laboratory buildings by Eddington. As they walked together down a corridor, they came to an open door. Chapman looked through it and ..... a large lecture theatre filled ..... students listening intently to the ..... at the front.

"What's going ..... in there?" asked Chapman.

"That's a meeting of mathematicians", replied Eddington. "..... I didn't know that there ..... so many mathematicians in the ..... world!" cried Chapman.

Eddington pulled ..... away from the door. "Between ..... and me", he answered,"there ....."

Eddington's moral, as I shall ..... it, is simple but important; ..... are not a lot of ..... in the world, but many ..... are going to mathematics lectures ..... taking degrees in mathematics or ..... related subject. Indeed, the moral ..... be extended; there are even ..... people studying mathematics in schools ..... institutes, and therefore an even ..... proportion of them possess substantial ..... ability.

Yet mathematics is not ..... for mathematicians; everybody requires a ..... amount of mathematical knowledge for ..... daily lives, and a perceptible ....., such as those who attend ..... lectures during their degree courses, ... considerably more. So the educational ..... are, basically; What should be ..... ? To whom? How? When? And ..... all, Why?

Start ..... Finish .....

CC 31

It is probable that in the foreseeable future a sex-control technology will be developed which will make it possible for parents to determine the sex of their next child. The use of this kind of technology will have important implications. In the following paragraph we discuss the possible effect on the composition of families.

What effect would the use of sex-control technology have on the sex and birth order of the offspring of individual couples? Estimates are presented ..... two- and three-child families. In order ..... estimate these effects for ..... two-child case, ... first calculated the probability that childless women ..... prefer the first child to ..... a male ..... a female; we ..... calculated, for women ..... one child, the conditional preference ..... the second child, given ..... sex ..... the first. The product of ..... two possibilities yields the expected proportion ..... conditions of sex pre-selection ..... each of ..... four possible birth order ..... sex permutations. ..... a similar manner the calculations ..... extended ..... the three-child case. The results ..... these calculations and the currently prevailing distributions ..... shown in Table 2 ..... the first two births and in Table 3 .... first three births.

Start .....

Finish .....

READING TEST PASSAGES - ORIGINALS

Passage AC

The most recent predictions strongly discount the possibility of seeing the comet in full daylight; they put its maximum brightness as close to that of Venus (which by the way is a brilliant object at present visible early in the evening as it sets in the south-west). Unfortunately our northern latitude does not favour the viewing of this particular comet, which rises and sets at a small angle to the horizon and, at its brightest period, appears very low in the sky.

The comet was photographed at the Observatory when its brightness was lower than that of the faintest naked eye star(s). It will, however, continue to get brighter as it approaches the sun, and by the middle of the/this month it will be readily visible to the naked eye in the morning sky, a few degrees above the horizon in a direction east/west of south about half an hour before sunrise. When at its most luminous, on the 28th, it will be only one degree away from the sun and, unless it becomes a brilliant daylight comet after all, it will be invisible to us.

Source: University Bulletin 10.5, Edinburgh, December 1973

Passage BC

The chapter on magic is especially important and valuable because of the abundance of texts, of the details given as to the ritual and the ingredients employed, and the inside information on the psychology of act and situation alike. Some anthropologists may question here the reinterpretation of the real processes which underlie magic. "It can safely be said that this (ie magic) is one way of transmitting thoughts telepathically from one mind to another," we are told. "It seems that through concentration, the magician or the possessor of love magic is able to penetrate into the mental mechanism of the person with whom he desires to establish communication. In this form the magician's suggestions are easily transmitted by means of vibration(s) to the brain, and then to the mind."

I submit that the author would still have to supply some evidence as to how these 'vibrations' are produced, how they act on the brain, and thence on the mind. But the author disarms all criticism when he tells us: "There is something in it which can be classified as occultism, and, as such, it cannot be dismissed as merely supersitition."

Source: B Malinowski in Introduction to Jomo Kenyatta, Facing Mt Kenya, London, Secker and Warburg, 1938.

Passage CC

When he first went to Cambridge as a young man, Chapman was shown round the laboratory buildings by Eddington. As they walked together down a corridor, they came to an open door. Chapman looked through it and saw a large lecture theatre filled with students listening intently to the lecturer at the front.

"What's going on in there?" asked Chapman.

"That's a meeting of mathematicians," replied Eddington. "But I didn't know that there were so many mathematicians in the whole world!" cried Chapman.

Eddington pulled him away from the door. "Between you and me," he answered, "there aren't."

'Eddington's moral' as I shall call it, is simple but important; there are not a lot of mathematicians in the world, but many people are going to mathematics lectures and taking degrees in mathematics or some related subject. Indeed, the moral may be extended; there are even more people studying mathematics in schools and institutes, and therefore an even smaller proportion of them possess substantial mathematical ability.

Yet mathematics is not only for mathematicians; everybody requires a certain amount of mathematical knowledge for their daily lives, and a perceptible minority, such as those who attend mathematics lectures during their degree courses, need considerably more. So the educational questions are, basically: What should be taught? To whom? How? When? And above all, Why?

Source: not traceable.

Passage MC

What effect would the use of sex-control technology have on the sex and birth order of the offspring of individual couples? Estimates are presented for two- and three-child families. In order to estimate these effects for the two-child case, we first calculated the probability that childless women would prefer the first child to be a male or a female; we then calculated, for women with one child, the conditional preference for the second child, given the sex of the first. The product of these two possibilities yields the expected proportion under conditions of sex pre-selection in each of the four possible birth order and sex permutations. In a similar manner the calculations were extended to the three-child case. The results of these calculations and the currently prevailing distributions are shown in Table 2 for the first two births and in Table 3 for the first three births.

Source: Journal of the American Association for the Advancement of Science, 10 May 1974

Passage DC

A great improvement was made in British farming between 1700 and 1800, when new scientific methods began to be introduced. The best known of the men who brought in new ideas was Lord Townsend, known as 'Turnip Townsend', who farmed in Norfolk between 1730 and 1738.

Lord Townsend introduced a way of using the land more profitably by growing in rotation turnips, barley, grass and wheat, thus avoiding the need to let land rest every third or fourth year. He also solved the problem of feeding cattle in winter. Previously, when grass stopped growing, farmers had to feed their cattle on hay and there was never enough, so most of the animals were killed, and the meat salted to prevent it going bad. Lord Townsend grew crops of turnips and stored them to feed the cattle in the winter, and so it was not necessary to slaughter so many.

The new methods of growing crops meant smaller fields, so hedges or walls divided up the old large fields, making the country look even more like it does now. The old borders were followed, however, and the shapes of today often mark the edges of fields and cultivation strips of a thousand years ago.

Source: R Bowood, Our Land in the Making - Book 2, Ladybird Books, Loughborough, Wills and Hepworth, 1966

APPENDIX II - FIRST INVESTIGATION - SUPPLEMENTARY BACKGROUND DATA

Section 1 Supplementary background data - whole sample

Note: The table numbers below refer to the relevant chapter and table number (first two digits).

Table 5.4.1 Countries of origin by geographic area

	Country	No		Country	No
	AFRICA			E and SE EUROPE	
1.	Cameroon	2	1.	Cyprus	2
2.	Congo	4	2.	Czechoslovakia	4
3.	Ethiopia	22	3.	Greece	10
4.	Malawi	3	4.	Poland	19
5.	Tanzania	3	5.	Turkey	39
6.	Zaire	1	6.	Yugoslavia	8
	Total	35		Total	82
	% of sample	4.2%		% of sample	9.8%

S and C AMERICA		
1.	Argentina	15
2.	Bolivia	3
3.	Brazil	40
4.	Chile	37
5.	Colombia	12
6.	Costa Rica	11
7.	Dominican Rep	2
8.	Ecuador	6
9.	Honduras	1
10.	Mexico	76
11.	Nicaragua	5
12.	Panama	1
13.	Peru	24
14.	Salvador	2
15.	Uruguay	6
16.	Venezuela	9
	Total	250
	% of sample	30%

MID EAST and N AFRICA		
1.	Algeria	25
2.	Bahrain	10
3.	Egypt	27
4.	Iran	29
5.	Iraq	11
6.	Israel	7
7.	Jordan	15
8.	Kuwait	4
9.	Lebanon	2
10.	Morocco	3
11.	Oman	4
12.	Saudi Arabia	2
13.	Sudan	49
14.	Tunisia	2
15.	Yemen	5
	Total	195
	% of sample	23.4%

S and E ASIA		
1.	Afghanistan	12
2.	Bangladesh	5
3.	Burma	27
4.	Indonesia	41
5.	Japan	16
6.	Khmer Rep	10
7.	Korea	10
8.	Laos	1
9.	Nepal	24
10.	Thailand	56
11.	Vietnam	4
	Total	206
	% of sample	24.7%

N and SW EUROPE		
1.	Austria	5
2.	Belgium	3
3.	Denmark	7
4.	Finland	2
5.	France	9
6.	Germany	15
7.	Iceland	3
8.	Italy	6
9.	Malta	1
10.	Netherlands	7
11.	Norway	4
12.	Portugal	1
13.	Spain	2
	Total	65
	% of sample	7.8%

Section 2 Supplementary background data by sub-sample

Table 5.2.1 Age

Age	EPTB %	BCSA %	Total sample %
18-20	2.3	0.3	1.3
21-25	15.7	18.9	17.2
26-30	42.6	41.9	42.3
31-35	26.5	27.1	26.8
36-40	8.5	9.2	8.9
41-45	3.5	2.1	2.8
46-50	0.7	0.5	0.6
51-55	0.2	-	0.1
Mean age (yrs)	29.96	29.79	29.9

Table 5.4.2 Countries of origin by geographic area

	AFRICA			E and SE EUROPE		
	Country	N(EPTB)	N(BCSA)	Country	N(EPTB)	N(BCSA)
1.	Cameroon	-	2	Cyprus	-	2
2.	Congo	-	4	Czechoslovakia	-	4
3.	Ethiopia	-	22	Greece	4	6
4.	Malawi	-	3	Poland	19	-
5.	Tanzania	3	-	Turkey	35	4
6.	Zaire	-	1	Yugoslavia	-	8
	Total	3	32	Total	58	24
	% of sample	0.4	3.8	% of sample	6.9	2.9
	% of regional sample	8.5	91.5	% of regional sample	70.7	29.3

S and C AMERICA

	Country	N(EPTB)	N(BCSA)
1.	Argentina	6	9
2.	Bolivia	-	3
3.	Brazil	1	39
4.	Chile	-	37
5.	Colombia	6	6
6.	Costa Rica	-	11
7.	Dominican Rep	-	2
8.	Ecuador	-	6
9.	Honduras	-	1
10.	Mexico	69	7
11.	Nicaragua	-	5
12.	Panama	-	1
13.	Peru	14	10
14.	Salvador	-	2
15.	Uruguay	-	6
16.	Venezuela	-	9
	Total	96	154
	% of sample	11.5	18.5
	% of regional sample	38.4	61.6

MID EAST and N AFRICA

	Country	N(EPTB)	N(BCSA)
1.	Algeria	21	4
2.	Bahrain	-	10
3.	Egypt	23	4
4.	Iran	23	6
5.	Iraq	9	2
6.	Israel	-	7
7.	Jordan	-	15
8.	Kuwait	-	4
9.	Lebanon	-	2
10.	Morocco	-	3
11.	Oman	-	4
12.	Saudi Arabia	-	2
13.	Sudan	41	8
14.	Tunisia	-	2
15.	Yemen	4	1
	Total	121	74
	% of sample	14.5	8.9
	% of regional sample	62.1	37.9

S and E ASIA

	Country	N(EPTB)	N(BCSA)
1.	Afghanistan	-	12
2.	Bangladesh	-	5
3.	Burma	26	1
4.	Indonesia	39	2
5.	Japan	16	-
6.	Khmer Rep	-	10
7.	Korea	-	10
8.	Laos	-	1
9.	Nepal	24	-
10.	Thailand	55	1
11.	Vietnam	-	4
	Total	160	46
	% of sample	13.2	5.5
	% of regional sample	77.7	22.3

N and SW EUROPE

	Country	N(EPTB)	N(BCSA)
1.	Austria	-	5
2.	Belgium	-	3
3.	Denmark	-	7
4.	Finland	-	2
5.	France	-	9
6.	Germany	-	15
7.	Iceland	-	3
8.	Italy	-	6
9.	Malta	-	1
10.	Netherlands	-	7
11.	Norway	-	4
12.	Portugal	-	1
13.	Spain	1	1
	Total	1	64
	% of sample	0.1	7.7
	% of regional sample	1.5	98.5

Table 5.5.1 Education completed

Education	EPTB %	BCSA %	Total %
1. Primary	-	-	-
2. Secondary	7.8	8.9	8.3
3. First degree or professional	63.5	64.0	63.7
4. Second degree (inc. MD)	22.1	22.7	22.4
5. Reserach degree or research	5.0	2.8	4.0
6. Diploma	1.6	1.5	1.6

Table 5.6.1 Subject area of study

Level	EPTB %	BCSA %	Total %
1. Agricultural, veterinary sciences	9.3	7.4	8.4
2. Arts	1.4	5.3	3.2
3. Medicine, nursing, psychology	10.7	11.2	10.9
4. Physical and biological sciences	16.6	12.2	14.5
5. Education, incl TEFL	8.9	15.0	11.8
6. Engineering and technology	18.5	15.0	16.8
7. English studies, (applied) linguistics	4.6	8.1	6.3
8. Business, social sciences (professional)	14.4	15.5	14.9
9. Social sciences (academic)	15.5	8.9	12.4
10. Miscellaneous	0.2	1.3	0.7

Table 5.7.1 Level of study

Level	EPTB %	BCSA %	Total %
2. Pre-university	0.2	0.8	0.5
3. First degree	6.6	2.5	4.7
4. Masters degree by tuition	30.5	20.6	25.8
5. Research, research degree	20.0	26.0	22.8
6. PG or professional Diploma/Cert	35.3	42.0	38.5
7. Academic attachment	4.2	5.0	5.0
8. Practical attachment	3.0	2.3	2.6

APPENDIX III CRITERION MEASURES - FINAL VERSIONS (used in the second investigation)

Section 1 English Ability Rating

- English Ability Rating form
- Key to English Ability Rating form
- Letter to British Council Regional Directors dated January 1975
- Letter to university tutors and supervisors dated January 1975

Student's Name ..... Field of Study .....

Country ....

ENGLISH ABILITY RATING

1. General ability in English
(Put X in the appropriate box in part A and in part B)

A

☐ completely adequate ☐ adequate ☐ only just adequate ☐ inadequate

B

☐ Shows native speaker ability in English usage and pronunciation

☐ Clearly a non-native speaker because of minor faults in English usage and pronunciation, but this does not handicap him/her in his/her studies

☐ Makes many mistakes in English, but this constitutes only a minor handicap for him/her in his/her studies

☐ Many weaknesses in English usage but his/her English ability can be considered just adequate for his/her studies. A higher standard is desirable

☐ Shows considerable deficiencies in English usage, which constitutes a handicap for him/her in his/her studies. A higher standard is necessary

☐ Shows very little ability in English and is well below a satisfactory standard

2. Individual language skills
(Put X at an appropriate point along the scale for each skill)

	completely adequate		adequate		only just adequate		inadequate
Ability to understand spoken English							
Ability to speak English							
Ability to understand ~~spoken~~ written English							
Ability to write English							

3. Improvement in English ability since beginning course of study

☐ considerable improvement ☐ a little improvement ☐ no improvement ☐ not applicable

4. Tuition in English since beginning course of study

☐ Yes ☐ No, don't know, or not applicable

5. Qualification aimed at, ... or

☐ on attachment only ☐ research only

6. Please use the back of this form for further comments

Date ..... Tutor .....

ENGLISH ABILITY RATING

Key for completing the accompanying English ability rating form

- Please complete one form per student
- The information requested below should be supplied in the boxes or spaces on the accompanying rating form

1. To what extent is the student's general ability in English now adequate for undertaking his/her specialised studies, research or training (i.e. as at Jan./Feb. 1975) ?

   Please answer this question by putting X first in the appropriate box in part A and then in the appropriate box in part B. The words 'and pronunciation' should be deleted (part B) if not applicable.

2. Please consider the student's ability in the following language skills –
   ability to understand spoken English
   ability to speak English
   ability to understand written English
   ability to write English
   Where possible give your rating of his/her ability in each skill by putting X at an appropriate point along the scale from completely adequate (i.e. native speaker ability) to inadequate (i.e. little or no ability).

3. Do you think the student has shown any improvement in his/her English ability since beginning the course ? Put X in the appropriate box. This question and the next may not be applicable for students who arrived with a very high standard of English.

4. Has the student received any tuition in English since beginning the course ? Put X in the appropriate box.

5. Please indicate which degree, diploma or other qualification the student is working for, or whether he/she is on attachment or doing research only. Put X in the appropriate box.

6. Please use the reverse of the rating form for any further comments you may wish to make.

- Please check that the student's field of study is correctly given at the head of the rating form and put the date and your own signature at the foot of the form.

1/75.

Telephone: 031 - 667 - 1011
Extension 6722

Department of Linguistics
University of Edinburgh
Adam Ferguson Building
George Square
Edinburgh EH8 9LL

January 1975

You may remember that last year you and your staff kindly helped me with the distribution of some documents to members of the University staff, and staff of other institutions in your area, in connection with my research into English language proficiency of overseas students under the supervision of Dr. Alan Davies at Edinburgh University.. (I am a Council officer doing further training at Edinburgh.) I should like to thank you very much for your help and let you know that the response from tutors was very good and most helpful to my work.

I am obliged to carry out the same exercise this year and should be very grateful if you could once again arrange for the forms to be delivered to the University and addressed to appropriate supervisors or Heads of Department for completion. Each supervisor or Head of Department should receive

- one form letter,

- one copy of the key for completing the forms, plus

- all the forms bearing the names of students whom he is supervising or for whom he is responsible.

I should also be grateful if the supervisor's name and department could be inserted at the top of the form letter. If there are any difficulties, please let me know.

I apologise for the inconvenience that this request may cause you, but with a project of this scale I am forced to request your help. I shall be most grateful for your co-operation.

Yours sincerely,

A. D. Moller.

Telephone: 031 - 667 - 1011
Extension 6722

Department of Linguistics
University of Edinburgh
Adam Ferguson Building
George Square
Edinburgh EH8 9LL

January 1975

English Language Ability of Overseas Students

Last year I asked several hundred tutors and supervisors for their assistance in a research project into the English language ability of students from overseas which I am undertaking under the supervision of Dr. Alan Davies of the Department of Linguistics at the University of Edinburgh and with the co-operation and support of the British Council. A brief questionnaire was sent to tutors for completion. More than 80% of those distributed were returned to me, often with useful comments, and the information obtained has been of great value. I should like to thank everybody who was good enough to complete and return the questionnaire.

I am obliged, however, to seek your further co-operation this year and am enclosing copies of the English Ability Rating form and of the key explaining how to complete the form. I should be most grateful if you, or the member of the departmental staff who has most contact with the students named, could complete the forms and return them to me at the above address by 28th February 1975.

The sample of students has been selected from those who are administered by the British Council and whose English was assessed before leaving their home country and who began their study within the last six months. The rating form has been designed for completion by supervisors who are not specialists in language teaching. If there is an item which you nevertheless feel unable to complete, please leave it blank and complete as much of the form as possible. All assessments will remain confidential. Neither the details of the individual assessments nor the identities of the assessors will be divulged to any other person or body. Part of my research this year will involve interviewing and testing some of the students, possibly including one or more of yours. I hope to do this during the current term and shall be grateful for any support you can give. I shall be contacting the students concerned directly.

It is hoped that the results of this research may lead to

1. more efficient procedures in the selection of overseas students for awards to universities and colleges in Britain

2. improvement in the assessment of candidates' English

3. more specific indications as to the type of remedial English work that some students will need.

I am sure you will agree that this is an important area of investigation and trust that you will be willing to spare a few minutes of your time. If there is any further information you require, please contact me.

I look forward to receiving the completed forms and thank you in advance for your help.

Yours sincerely,

A. D. Moller.

Section 2 Communicative Proficiency Measure

- Reading Test instructions
  - Passage AC
  - Passage CC
  - Passage DC
  - Passage MC

- Writing Test instructions
  - Assessment scale

- Interview Test contents and phases
  - Assessment scale

- Letter to students

NAME ..... COLLEGE UNIVERSITY .....

COUNTRY ..... DEPARTMENT .....

READING

Here are four passages taken from recent books or journals.

In each passage you will find some blanks (.....). Each blank indicates one missing word.

Read as much of each passage as you can in 30 minutes.

Write in each blank a word which you think best fits that blank.

Fill in as many blanks as you can. You may attempt the passages in any order.
NB Each blank stands for one word only. Contractions, eg "can't", "won't", count as one word.

EXAMPLE

If costs go ..... any higher, the situation will ..... much more serious than it ..... today.

(The missing words are "up" "become" "is")

This passage is about a comet that has recently been discovered. The comet is noteworthy for its brightness and its comparative closeness to the sun. There is a lot of speculation as to whether we shall be able to see it in daylight later this month.

The most recent predictions strongly discount the possibility of seeing the ..... in full daylight; they put ..... maximum brightness at close to ..... of Venus (which by the ..... is a brilliant object at ....., visible early in the evening ..... it sets in the south-west). ..... our northern latitude does not ..... the viewing of this particular ..... , which rises and sets at ..... small angle to the horizon ..... , at its brightest period, appears ..... low in the sky.

The .....was photographed at the Observatory ..... its brightness was lower than ..... of the faintest naked eye ..... . It will, however, continue to ..... brighter as it approaches the ..... , and by the middle of ..... month it will be readily .....to the naked eye in ..... morning sky, a few degrees ..... the horizon in a direction ..... of south about half an ..... before sunrise. When at its ..... luminous, on the 28th, it .....be only one degree away ..... the sun, and unless it ..... a brilliant daylight comet after ..... , it will be invisible to ..... .

When he first went to Cambridge as a young man, Chapman was shown round the university laboratory buildings by Eddington. As they walked together down a corridor, they came to an open door. Chapman looked through it and ..... a large lecture theatre filled ..... students listening intently to the..... at the front.

"What's going ..... in there?" asked Chapman.
"That's ..... meeting of mathematicians", replied Eddington.
"..... I didn't know that there ..... so many mathematicians in the ..... world!"cried Chapman.

Eddington pulled ..... away from the door. "Between ..... and me", he answered, "there ..... !"

Eddington's moral , as I shall call it, is simple but important; ..... are not a lot of ..... in the world, but many ..... are going to mathematics lectures ..... taking degrees in mathematics or ..... related subject. Indeed, the moral ..... be extended; there are even ..... people studying mathematics in schools ..... institutes, and therefore, an even ..... proportion of them possess substantial ..... ability.

Yet mathematics is not ..... for mathematicians; everybody requires a ..... amount of mathematical knowledge for .. .. daily lives, and a perceptible ..... , such as those who attend ..... lectures during their degree courses, ... considerably more. So the educational ..... are, basically; What should be .....? To whom? How? When? And ..... all, Why?

A great improvement was made in British farming between 1700 and 1800, when new scientific methods began to be introduced. The best known of the men who brought in new ideas was Lord Townsend, known as 'Turnip Townsend', who farmed in Norfolk between 1730 and 1738.

Lord Townsend introduced a way ..... using the land more profitably ....., growing in rotation turnips, barley, ... and wheat, thus avoiding the ..... to let land rest every ..... or fourth year. He also ..... the problem of feeding cattle ..... winter. Previously, when grass stopped ..... , farmers had to feed their ..... on hay and there was ..... enough, so most of the ..... were killed, and the meat ..... to prevent it going bad. ..... Townsend grew crops of turnips ..... stored them to feed the ..... in the winter, and so ..... was not necessary to slaughter ..... many.

The new methods of ..... crops meant smaller fields, so ..... or walls divided up the ..... large fields, making the country ..... even more like it does ..... The old borders were followed, ... , and the field shapes of ..... often mark the edges of ..... and cultivation strips of a thousand years ago.

dc 25

It is probable that in the foreseeable future a sex-control technology will be developed which will make it possible for parents to determine the sex of their next child. The use of this kind of technology will have important implications. In the following paragraph we discuss the possible effect on the composition of families.

What effect would the use of sex-control technology have on the sex and birth order of the offspring of individual couples? Estimates are presented ..... two- and three-child families. In order ..... estimate these effects for ..... two child case, ... first calculated the probability that childless women ..... prefer the first child to ..... a male ..... a female; we ..... calculated for women ..... one child, the conditional preference ..... the second child, given ..... sex ..... the first. The product of ..... two possibilities yields the expected proportion ..... conditions of sex pre-selection ..... each of ..... four possible birth order and sex permutations. ..... a similar manner the calculations ..... extended ....., the three-child case. The results ..... theses calculations and the currently prevailing distributions ..... shown in Table 2 ..... the first two births and in Table 3 .... first three births.

WRITING

(After living abroad for many years I have returned to Britain to study for a higher degree and am finding the experience very rewarding. I hope that you are finding your stay as valuable. I should like to learn something about your stay in Britain.

Please write four or five paragraphs about one aspect of your work or life in Britain that interests, or has interested, you very much. Describe this work or experience and explain why you have found it interesting. To what extent has it been of value to you?

Notes for your guidance:

- You should be fairly brief as you have only 30 minutes. Leave yourself enough time to give some reasons for your interest in the work or experience.

- Accuracy of fact is not as important as fluency and style. You may include both favourable and adverse criticisms!

- I am looking for information on what you have done, or what you are doing, and I am looking for opinions.

Assessment Scale

Level 6. Near native speaker ability (with or without a few minor faults).

5. Some minor faults (probably in style, choice of lexis or occasional points of grammar), but fluent and generally very competent in communication of information and ideas.

4. Competent in communication of information and ideas, but displaying some weaknesses in style or use of vocabulary or grammatical accuracy. Not always very fluent but adequate.

3. Serious weaknesses are evident, and at times the successful communication of information or ideas is threatened. Barely adequate for academic writing.

2. Serious deficiencies in the control of grammar, sentence structure and lexis limit the effective communication of information or ideas. Inadequate.

1. Little or no ability to communicate effectively in written English.

INTERVIEW TEST

Content and phases

Phase 1. Introductory

- statement by interviewer
- setting at ease with checking of some administrative and personal details
- transition to next phase by asking about family or mail service or personal welfare.

Phase 2. Personal background

- home country
- travel to Britain
- accommodation - discussion
- life in Britain - discussion

Phase 3. Work being undertaken

- details of course
- description of typical day
- description of project/research - discussion
- discussion of value on return to home country

Phase 4. English language ability

- assessment in home country
- courses followed at home or in Britain - discussion
- problems encountered
- self-assessment and improvement - discussion

Phase 5. Questions

- any questions that subject may wish to put and answers/discussion from interviewer

Phase 6. Closing the interview

- choice of moment to close
- expression of thanks by the interviewer.

Assessment scale

Level 6. Native speaker ability and fluency with (occasional) evidence of non-native speaker pronunciation features.

5. Full communication established, but with occasional minor faults of English usage and pronunciation.

4. Adequate communication, occasionally impaired by a number of minor faults of usage or pronunciation. There may be occasional hesitations.

3. Communication established but impaired from time to time as a result of frequent faults in grammar/ lexical choice/pronunciation. There may be hesitations and expression may come in short or disjointed phrases.

2. Frequent limitations in communication by serious deficiencies in grammatical control and pronunciation.

1. Hardly any or no ability to communicate.

Department of Linguistics
Adam Ferguson Building,
George Square,
Edinburgh EH8 9LL

Dear

I am a postgraduate student doing research into English language proficiency testing under the supervision of Dr. Alan Davies at the University of Edinburgh. Part of my work involves assessing the English proficiency of a sample of students from overseas whose English was assessed before they came to study in Britain. I am sending a questionnaire to supervisors of students like yourself asking them to make an assessment of the English proficiency of their students from overseas. I am also anxious to interview and administer a short English test to the same students, and am asking you if you would be willing to be tested by me when I visit your institution or the British Council centre in

.....

The 'test' would take between 60 and 75 minutes and you could start at any time. There would be no preparation to do and I would be asking you to write and talk briefly about your work here, your work in your home country, your impressions of living in Britain and any language problems you may have had. I should stress that the results are of interest to me for my research and would be sent to you, if you wished. They would not be communicated to your supervisors or any other person or body.

As you have probably guessed, my work is being carried out with the support of the British Council, and it is through them that I have obtained your name.

As a fellow student I appreciate that you have very little time to spare but I should be very grateful for your co-operation. I am looking for students of <u>any</u> level of proficiency, so please don't think your English is too good or too bad to be assessed ! I hope, too, that the results of my research might lead to better testing procedures before admission to British universities, colleges, etc., and to better programmes of remedial English. I hope very much to see you on

..... in .....

.....

you may arrive at any time from ..... to .....

I should be grateful if you could complete and return the attached slip of paper to

Yours sincerely,

Alan Moller.

.....

To: .....

.....

From: .....

.....

I <u>shall</u> * be able to come to the English test
<u>shall not</u>
arranged by Mr. Alan Moller on .....

I shall be there at about .....

* Delete where not applicable.

APPENDIX IV - SECOND INVESTIGATION: DATA AND RESULTS

(Note: figure and table numbers are preceded by the chapter number to which they relate.)

Section 1 Background data of the whole sample

Figure 7.1 Personal Data and Test Results Sheet

Name	....	Name of subject
Sex	....	Sex
Age	....	Age
Coun	....	Country
Qual	....	Highest educational qualification
T1/L	....	Test 1 score/Listening grade
T2/S	....	Test 2 score/Speaking grade
T3/R	....	Test 3 score/Reading grade
T4/W	....	Test 4 score/Writing grade
Tot	....	EPTB Pt 1 Total/BCSA 'total'
5	....	Test 5 Score
O	....	Oral grade/as T2/S
W	....	Oral grade/as T4/W
Rem E	....	Weeks of remedial English
Univ	....	British institution attended
Dept/Subj	....	Subject area
Level	....	Level at which studying
Length	....	Length of course

Table 7.2 Distribution by age

Age	Frequency	Frequency %
18-20	9	1.2
21-25	129	17.9
26-30	280	38.8
31-35	188	26.0
36-40	69	9.6
41-45	36	5.0
45-50	11	1.5
	722	100.0

missing cases 7

mean 30.3 yrs (1st investigation 29.9 yrs)

mode 27 yrs (1st investigation 27 yrs)

Table 7.3 Distribution of sample by geographic area

No of countries	Areas	Frequency	Frequency %	1st investn frequency %
14	Africa (South of the Sahara)	80	11.0	4.2
14	Middle East & North Africa (inc Sudan)	178	24.4	23.4
16	South, Southeast and East Asia	165	22.6	24.7
18	Latin and Central America	203	27.9	30.0
14	North and Southwest Europe	48	6.6	7.8
6	East and Southeast Europe	55	7.5	9.8
82		729	100.0	100.0

Table 7.4 Countries of origin by geographic area

Country	No	Country	No
AFRICA		MID EAST and N AFRICA	
1. Botswana	11	1. Abu Dhabi	2
2. Cameroon	1	2. Algeria	29
3. Ethiopia	23	3. Bahrein	11
4. Kenya	1	4. Egypt	20
5. Lesotho	4	5. Iran	15
6. Malawi	3	6. Iraq	7
7. Mali	1	7. Israel	5
8. Niger	1	8. Jordan	19
9. Senegal	2	9. Kuwait	3
10. Sierra Leone	3	10. Lebanon	1
11. Swaziland	2	11. Morocco	3
12. Tanzania	7	12. Oman	5
13. Togo	1	13. Sudan	49
14. Zambia	20	14. Yemen	9
Total	80	Total	178
% of sample	11.0	% of sample	24.4
S and C AMERICA		S and E ASIA	
1. Argentina	16	1. Afghanistan	11
2. Bolivia	6	2. Bangladesh	5
3. Brazil	36	3. Burma	9
4. Chile	17	4. Indonesia	31
5. Colombia	13	5. Japan	22
6. Costa Rica	5	6. Khmer Rep	3
7. Dominican Rep	2	7. Korea	10
8. Ecuador	2	8. Laos	5
9. Haiti	1	9. Nepal	10
10. Honduras	3	10. New Hebrides	5
11. Mexico	55	11. Pakistan	6
12. Nicaragua	3	12. Solomon Is	6
13. Panama	6	13. Sri Lanka	2
14. Paraguay	1	14. Thailand	31
15. Peru	24	15. Tonga	1
16. Salvador	6	16. Vietnam	8
17. Uruguay	1		
18. Venezuela	6		
Total	203	Total	165
% of sample	27.9	% of sample	22.6

Country	No	Country	No
N and SW EUROPE		E and SE EUROPE	
1. Austria	2	1. Cyprus	6
2. Belgium	4	2. Czechoslovakia	4
3. Denmark	1	3. Greece	4
4. Finland	1	4. Poland	11
5. France	8	5. Turkey	23
6. Germany	12	6. Yugoslavia	7
7. Iceland	1		
8. Italy	4		
9. Malta	2		
10. Netherlands	3		
11. Portugal	3		
12. Spain	3		
13. Sweden	1		
14. Switzerland	3		
Total	48	Total	55
% of sample	6.6	% of sample	7.5

Table 7.5 Distribution of largest country groups

Country	Frequency	% of total sample	1st investn Frequency %
Mexico	55	7.5	9.1
Sudan	49	6.7	5.9
Brazil	36	4.9	4.8
Algeria	29	4.0	3.0
Thailand	31	4.3	6.7
Indonesia	31	4.3	4.8
Peru	24	3.3	2.9
Ethiopia	23	3.2	2.6
Turkey	23	3.2	4.7
Japan	22	3.0	2.6
	323	44.4	47.1

Table 7.6 Educational background

	Education	Frequency	%	1st investn Frequency %
1.	Primary only	-	-	-
2.	Secondary or equivalent	194	21.0	8.3
3.	First degree or professional qualification	544	58.8	63.7
4.	Second degree (inc MD)	147	15.9	22.4
5.	Research degree or research	36	3.9	4.0
6.	Diploma	4	0.4	1.6
		25	100.0	100.0

Table 7.7 Distribution of sample by area and institution in Britain

Area/institution	Frequency	% of sample
Edinburgh incl Moray House	58	8.0
Manchester and Salford	55	7.5
Leeds	39	5.3
Oxford	38	5.2
Cambridge	34	4.7
Reading	30	4.1
Aston and Birmingham	28	3.8
Newcastle	26	3.6
Glasgow, incl Strathclyde	24	3.3
Imperial College, London	23	3.2
Totals	355	48.7

Table 7.8 Distribution of sample according to subject area of study

Subject area	Frequency	%	% change
1. Agriculture, veterinary sciences	61	8.4	same
2. Arts subjects	21	2.9	- 0.3
3. Medicine, nursing, psychology	53	7.3	- 3.6
4. Physical and biological sciences	83	11.4	- 3.1
5. Education, incl TEFL	118	16.2	+ 4.4
6. Engineering and technology	159	21.8	+ 5.0
7. English studies and (applied) linguistics	45	6.2	same
8. Business, social sciences (professional)	131	18.0	+ 3.1
9. Social sciences (academic)	53	7.3	- 5.1
10. Miscellaneous	4	0.5	- 0.2
Total	729	100.0	

(Details of first investigation are in Table 5.6)

Table 7.9 Distribution of sample according to level of study

		Frequency	%	1st investn frequency %
1.	-	-	-	-
2.	Pre-university	9	1.2	0.5
3.	First degree	40	5.5	4.7
4.	Master's degree by tuition	170	23.3	25.8
5.	Research, research degree	117	16.0	22.8
6.	PG or professional diploma/ certificate	303	41.6	38.5
7.	Academic attachment	69	9.5	5.0
8.	Practical attachment	21	2.9	2.6
		729	100.0	100.0

Section 2 Pre-departure assessment results - EPTB and BCSA sub-samples

Table 7.10 EPTB: Obtained means and SD's

	Sample Mean	Sample SD		Standard Mean	Standard SD
1. Phonemic discrimination	9.09	1.4	(N = 281)	10.0	2.0
2. Intonation	9.96	1.6	(N = 281)	10.0	2.0
3. Reading comprehension	9.6	1.5	(N = 286)	10.0	2.0
4. Grammar	9.33	1.7	(N = 286)	10.0	2.0
Part 1. Total	38.42	4.5	(N = 311)	40.0	6.0
Part 2. Speed reading	65.96	28.6	(N = 130)	70.0	34.0
				or 62.0	27.0

(Results for the first investigation sample are reported in Table 5.8 above.)

Table 7.11 EPTB: Part 1 total scores grouped according to interpretation

Part 1 Total	Interpretation	N	%	1st investn %
Below 34.0	English inadequate	53	16.9	11.2
34.0-39.9	Preliminary tuition needed	158	50.5	51.0
40.0-45.9	English should be adequate	81	25.8	31.2
46.0 and above	Unquestionably adequate	21	6.7	6.6
		313	100.0	100.0

Table 7.12 EPTB: Grades awarded for oral and writing tests

Grade	Oral N	Oral %	Writing N	Writing %
D-	-	-	3	1.9
D	1	0.6	1	0.7
D+	2	1.2	-	-
C-	4	2.4	2	1.3
C	8	4.7	35	23.2
C+	7	4.1	8	5.3
B-	7	4.1	8	5.3
B	63	37.3	51	33.8
B+	11	6.5	5	3.3
A-	4	2.4	1	0.7
A	62	36.7	37	24.5
Total	229	100.0	203	100.0
Mean	B+		B	

(First investigation details are given in Table 5.10.)

Table 7.13 BCSA: Summary of grades awarded

Gr	Listening N	Listening %	Speaking N	Speaking %	Reading N	Reading %	Writing N	Writing %	Total N	Total %
D	1	0.3	3	0.9	1	0.2	7	1.6	1	0.2
D+	-	-	-	-	1	0.2	1	0.2	3	0.6
C	19	5.2	49	10.7	34	7.7	85	19.3	23	5.0
C+	6	1.6	20	4.3	29	6.6	36	8.2	73	15.9
B	92	25.1	163	35.4	142	32.3	159	36.1	109	23.7
B+	42	11.4	41	8.9	17	3.9	15	3.4	142	30.9
A	207	56.4	183	39.8	216	49.1	137	31.1	109	23.7
	367	100.0	460	100.0	440	100.0	440	100.0	460	100.0
Mean	A-		B+		B+		B		B+	

(First investigation details are given in Table 5.11.)

Table 7.14 BCSA: First and second sample distributions of 'total' grades compared

Grade	First sample N	First sample %	Second sample N	Second sample %
D	1	0.3	-	-
D+	-	-	2	0.5
C	11	3.0	18	4.9
C+	41	10.9	26	7.0
B	66	17.6	124	33.4
B+	134	35.7	68	18.4
A	122	32.5	133	35.8
Total	375	100.0	460	100.0

Table 7.15 EPTB and BCSA: Comparison of grades for oral and writing

	Writing				Oral			
	EPTB		BCSA		EPTB		BCSA	
Grade	N	%	N	%	N	%	N	%
D	4	(2.7)	4	(1.1)	1	(0.6)	3	(0.8)
D+	2	(1.3)	1	(0.3)	6	(3.6)	-	(-)
C	35	(23.2)	66	(18.9)	8	(4.7)	38	(10.4)
C+	16	(10.6)	30	(8.6)	14	(8.2)	14	(3.8)
B	51	(33.8)	127	(36.3)	63	(37.3)	123	(33.5)
B+	6	(4.0)	13	(3.8)	15	(8.9)	38	(10.4)
A	37	(24.5)	109	(31.1)	62	(36.7)	151	(41.1)
	151	(100.0)	350	(100.0)	169	(100.0)	367	(100.0)
Mean	B (low)		B		B+ (low)		B+ (low)	

Table 7.16 Summary of numbers taking remedial English

	N	% sub-sample	% total
EPTB + remedial English	208	65.8	28.5
BCSA + remedial English	252	61.0	34.6
Total + remedial English	460	-	63.1
No remedial English	269	-	36.9
	729		100.0

(First investigation details are given in Table 5.13.)

Table 7.17 Length of remedial English tuition

	EPTB		BCSA		Total	
Weeks	N	%	N	%	N	%
2-4	65	31.2	75	29.8	140	30.4
5-8	66	31.7	77	30.6	143	31.1
9-12	72	34.6	79	31.3	151	32.8
16-25	5	2.5	21	8.3	26	5.7
	208	100.0	252	100.0	460	100.0

(First investigation details are given in Table 5.14.)

Section 3 Background data of the EPTB and BCSA sub-samples

Table 7.18 Distribution of age

Age	EPTB Freq	EPTB Freq %	BCSA Freq	BCSA Freq %	Whole Freq %
18-20	5	1.6	4	1.0	1.2
21-25	57	18.5	72	17.4	17.9
26-30	113	36.5	167	40.4	38.8
31-35	78	25.3	110	26.7	26.0
36-40	32	10.3	37	8.9	9.6
41-45	18	5.9	18	4.4	5.0
46-50	6	1.9	5	1.2	1.5
	309	100.0	413	100.0	100.0
Mean age	30.4 years		30.2 years		

Table 7.19

(a) EPTB sub-sample: distribution by country of origin

Country	Freq	Freq %	Country	Freq	Freq %
Algeria	14	4.4	Lesotho	2	0.6
Argentina	3	0.9	Mali	1	0.3
Bangladesh	2	0.6	Mexico	46	14.6
Botswana	8	2.5	Nepal	10	3.2
Brazil	1	0.3	Pakistan	1	0.3
Burma	4	1.3	Peru	17	5.4
Colombia	8	2.5	Poland	11	3.5
Egypt	15	4.7	Spain	1	0.3
Greece	4	1.3	Sudan	36	11.4
Honduras	1	0.3	Tanzania	7	2.2
Indonesia	28	8.9	Thailand	30	9.5
Iran	15	4.7	Turkey	21	6.6
Iraq	7	2.2	Vietnam	1	0.3
Japan	19	6.0	Yemen	3	0.9
			Total	316	100.0

(b) BCSA sub-sample: distribution by country of origin

Country	Freq	Freq %	Country	Freq	Freq %
Abu Dhabi	2	0.5	Laos	5	1.2
Afghanistan	11	2.7	Lebanon	1	0.2
Algeria	15	3.6	Lesotho	2	0.5
Argentina	13	3.1	Malawi	3	0.7
Austria	2	0.5	Malta	2	0.5
Bahrein	11	2.7	Mexico	9	2.2
Bangladesh	3	0.7	Morocco	3	0.7
Belgium	4	1.0	Netherlands	3	0.7
Bolivia	6	1.5	New Hebrides	5	1.2
Botswana	3	0.7	Nicaragua	3	0.7
Brazil	35	8.5	Niger	1	0.2
Burma	5	1.2	Oman	5	1.2
Cameroon	1	0.2	Pakistan	5	1.2
Chile	17	4.1	Panama	6	1.5
Colombia	5	1.2	Paraguay	1	0.2
Costa Rica	5	1.2	Peru	7	1.7
Cyprus	6	1.5	Portugal	3	0.7
Czechoslovakia	4	1.0	Salvador	6	1.5
Denmark	1	0.2	Senegal	2	0.5
Dominica	2	0.4	Sierra Leone	3	0.7
Ecuador	2	0.5	Solomon Is	6	1.5
Egypt	5	1.2	Spain	2	0.5
Ethiopia	23	5.6	Sri Lanka	2	0.5
Finland	1	0.2	Sudan	13	3.1
France	8	1.9	Swaziland	2	0.5
Germany	12	2.9	Sweden	1	0.2
Haiti	1	0.2	Switzerland	3	0.7
Honduras	2	0.5	Thailand	1	0.2
Iceland	1	0.2	Togo	1	0.2
Indonesia	3	0.7	Tonga	1	0.2
Israel	5	1.2	Turkey	2	0.5
Italy	4	1.0	Uruguay	1	0.2
Japan	3	0.7	Venezuela	6	1.5
Jordan	19	4.6	Vietnam	7	1.7
Kenya	1	0.2	Yemen	6	1.5
Khmer	3	0.7	Yugoslavia	7	1.7
Korea	10	2.4	Zambia	20	4.8
Kuwait	3	0.7			
			Total	413	100.0

Table 7.20 Distribution of sub-samples by geographic area

Area	EPTB Freq	EPTB %	BCSA Freq	BCSA %
Africa (South of the Sahara)	18	5.6	75	18.2
Middle East and N Africa (incl Sudan)	90	28.5	75	18.2
South, Southeast and East Asia	95	30.1	70	16.9
Latin and Central America	76	24.1	127	30.7
North and Southwest Europe	1	0.3	47	11.4
East and Southeast Europe	36	11.3	19	4.6
	316	100.0	413	100.0

Table 7.21 Educational background

Education	EPTB Freq	EPTB %	BCSA Freq	BCSA %
1. Primary only	-	-	-	-
2. Secondary or equivalent	53	16.8	97	23.5
3. First degree or professional qualification	201	63.6	234	56.7
4. Second degree (incl MD)	49	15.5	68	16.5
5. Research degree or research	12	3.8	13	3.1
6. Diploma	1	0.3	1	0.2
	316	100.0	413	100.0

Table 7.22 Subject areas of study

Subject area	EPTB Freq	EPTB %	BCSA Freq	BCSA %
Agriculture, veterinary sciences	25	7.9	36	8.7
Arts subjects	6	1.9	15	3.6
Medicine, nursing, psychology	19	6.0	34	8.2
Pure sciences	34	10.8	49	11.9
Education incl TEFL	40	12.7	78	18.9
Engineering and technology	84	26.6	75	18.2
English & linguistics	10	3.2	35	8.5
Social sciences (professional)	68	21.5	63	15.3
Social sciences (academic)	27	8.5	26	6.3
Miscellaneous	3	0.9	2	0.5
	316	100.0	413	100.0

Table 7.23 Level of study

Level	EPTB Freq	EPTB %	BCSA Freq	BCSA %
2. Pre-university	6	1.9	3	0.7
3. First degree	17	5.4	23	5.6
4. Master's degree by tuition	88	27.8	82	19.9
5. Research or research degree	46	14.6	71	17.2
6. PG or professional diploma/ certificate	119	37.7	184	44.6
7. Academic attachment	28	8.9	41	9.9
8. Practical attachment	12	3.8	9	2.2
	316	100.0	413	100.0

Section 4 English Ability Ratings by tutors - results

Table 7.24 Tutors' ratings - question 1A

Category	Frequency	%
1. Inadequate English	16	2.2
2. Only just adequate	157	21.6
3. Adequate	372	51.2
4. Completely adequate	181	24.9
	726	100.0

Table 7.25 Tutors' ratings - question 1A - EPTB and BCSA sub-samples

Category	EPTB Frequency %	BCSA Frequency %
1. Inadequate English	2.8	1.7
2. Only just adequate	26.6	17.8
3. Adequate	53.5	49.5
4. Completely adequate	17.1	31.0
	100.0	100.0
N=	316	410

Table 7.26 Tutors' ratings - question 1B

Category	Frequency	%
1. Totally inadequate	1	0.1
2. More profy essential	43	5.9
3. More profy desirable	132	18.1
4. Minor deficiencies	159	21.8
5. Adequate	351	48.2
6. Near native-speaker	42	5.8
	728	100.0

Table 7.27 Tutors ratings - question 1B - EPTB and BCSA sub-samples

Category	EPTB Frequency %	BCSA Frequency %
1. Totally inadequate	-	0.2
2. More profy essential	7.9	4.4
3. More profy desirable	22.8	14.6
4. Minor deficiencies	23.1	20.9
5. Adequate	41.5	53.4
6. Near native speaker	4.7	6.6
	100.0	100.0
N =	316	412

Table 7.29 Distribution of language skills ratings

Scale		Listen	Speak	Read	Write
1-4	inadequate	7	11	3	29
4.5-7	only just adequate	91	154	59	160
7.5-10	adequate	364	392	334	359
10.5-13	completely adequate	266	172	316	143
		728	729	712	691
	Mean rating	9.6	8.9	10.0	8.6
	SD	2.1	2.2	1.9	2.4

Comparable information on the first investigation sample is at Table 5.19 although different intervals were used.

Figure 7.30 Question 2, skills rating scale - summary of responses

	completely adequate	adequate	only just adequate	inadequate
Ability to understand spoken English	266	364	91	7
Ability to speak English	172	392	154	11
Ability to understand written English	316	334	59	3
Ability to write English	143	359	160	29

Table 7.31 Language skills ratings - intercorrelations

	Listen	Speak	Read	Write
Listening	-			
Speaking	.7192	-		
Reading	.6800	.6284	-	
Writing	.5929	.6700	.6655	-

(All coefficients beyond the 1% level of significance)

Table 7.33 Language skills ratings - EPTB and BCSA sub-samples: % frequencies

Scale		Listening		Speaking		Reading		Writing	
		EPTB	BCSA	EPTB	BCSA	EPTB	BCSA	EPTB	BCSA
1-4	inadequate	1.3	0.7	1.9	1.2	0.7	0.2	5.4	3.3
4.5-7	only just adequate	17.1	9.0	27.5	16.2	10.1	6.9	28.9	18.8
7.5-10	adequate	55.7	44.7	53.5	53.3	52.6	41.2	47.5	54.3
10.5-13	completely adequate	25.9	45.6	17.1	29.3	36.6	51.7	18.2	23.6
	Total N	316	412	316	413	306	406	297	394
	Mean rating	9.1	10.0	8.5	9.3	9.7	10.3	8.3	8.9
	SD	2.1	2.0	2.1	2.1	1.9	1.9	2.5	2.3

Table 7.34 Language improvement rated by tutors

Category	N	%	First sample %
1. Considerable	147	20.2	32.9
2. A little	407	56.0	62.6
3. None	82	11.3	4.5
4. Not applicable	91	12.5	-
	727	100.0	100.0

Table 7.35 Language improvement by sub-sample

Category	EPTB Freq	EPTB %	BCSA Freq	BCSA %
1. Considerable	68	21.5	79	19.2
2. A little	184	58.2	223	54.3
3. None	34	10.8	49	11.9
4. Not applicable	30	9.5	60	14.6
	316	100.0	411	100.0

Table 7.36 Language tuition since beginning course of study

	Whole		EPTB	BCSA
	Frequency	%	Frequency %	Frequency %
Yes	184	25.2	26.3	24.5
No/not known	545	74.8	73.7	75.1

Table 7.37 Correlations between tutors' ratings

(R1A represents section A of question 1, R1B represents section B of question 1, and RT represents grouped skills ratings in question 2).

	R1A	R1B	RT
R1A	-	.78 (N = 725)	.85 (N = 686)
R1B		-	.78 (N = 688)
RT			-

Pearson product moment correlations - all beyond 1% level of significance.

Table 7.38 Tutors' skills ratings - intercorrelations

RL = Listening RS = speaking RR = reading RW = writing
RT = Total

	RL	RS	RR	RW
RL	-	.72	.68	.59
RS	.72	-	.63	.67
RR	.68	.63	-	.67
RW	.59	.67	.67	-
RT	.86	.87	.85	.86

All correlations were significant at the 1% level. N ranged from 689 to 728.

Section 5 Comments made by tutors on the ratings form.

(Note: Numbers indicate the subject's identification number. Subjects with numbers between 1001 and 2070 are female)

1 (Iran)

So far, this student's performance has been only just satisfactory and it is difficult to assess whether this is primarily due to language difficulties or not.

11 (Colombia)

This student's performance so far has not been up to standard, largely due to his inability to express himself at all well in writing in the English language. He has been advised to continue with classes in English and he will be given ample opportunity in the next 3 months to answer sample examination questions etc. If he continues to improve his language ability at the same rate, it is just possible that he will eventually pull through.

13 (Nepal)

This student is so far doing very well on the MSc course, but is evidently requiring more time for background reading in this subject than most English students.

14 (Nepal)

Difficulty is experienced with Mr P on the subject with which he is already acquainted.

There is extreme difficulty in trying to teach him anything new and one is not sure how much he has absorbed. Further difficulties arise from differences in terminology, eg cable vault for cable chamber, splicing for jointing.

The main difficulty is that he himself can put only 2 or 3 words together when he tries to speak English.

1016 (Thailand)

Miss S is spending this year working towards a qualifying examination. If she passes this, then she will be allowed to proceed to the MSc course in Geophysics next year. As you will see from my comments her ability in English is a severe impediment to her progress, a feature we have grown to associate with British Council students. I was pleased to see today that Imperial College is to practise as routine the instruction for foreign language applicants to take the TOEFL examination.

1020 (Thailand)

She attended a language school before beginning the course at Sussex and my impression is that the language course was very successful. Science graduate students require facility in writing English in the past tense and preferably in the passive voice. This appears to cause difficulty for those whose mother tonge is one of the oriental languages.

24 (Turkey)

Somewhat shy and retiring. Does not use English sufficiently and so improves very slowly. Does not mix much with other students.

37 (Mexico)

When I compare the standard of English of the two students whom I have had or will have from Mexico (where I supervised their research before they came to the UK) with the results of the tests carried out by the British Council Assessor of linguistic ability there, I wonder if he/she and I are thinking about the same language. The standard used there is seemingly very high and in each case has caused the student an amount of concern out of proportion to the deficiencies in the students' written and oral abilities.

1040 (Mexico)

I should like to know whether this lady writes fluently in Spanish! Her answers to test questions, and her prepared essays, are very short and do not by any means do justice to her knowledge, which conversation reveals as more than adequate.

46, 48, 49, 56 (Sudan)

Questions very difficult to answer with any confidence at this stage of the course.

He has been attending lecture courses, and I feel test will come with examinations and research project work. We shall be in a better position to answer questions in 3 months' time. We can only judge him so far on conversations and with questions during classes. He may have difficulties which we do not yet appreciate.

58 (Tanzania)

This student has a good understanding of the English language and speaks it well. He himself, however, believes that his ability to write English has improved since he came on this course.

62 (Thailand)

Because the structure of sentences in English is vastly different from Thai language structures many Thai students experience difficulty in spoken English.

75 (Colombia)

1. I regret I cannot return the form by 28.2.75 because I have only just received it (28 Feb).

2. Question 1B. There are a few very minor faults but whether or not he finds these a handicap I do not know.

3. I am not sure what is implied by 'on attachment' in Question 5. As far as I am concerned Mr S is a full-time student on a course of lectures and practical work.

4. Whilst he is not perfect in English Mr S is considerably better than all the other students that have been put forward to me by the British Council. The other Council-recommended student I have on the course has still not reached the 'adequate' stage after six months in this country, receiving regular specialist tuition in English, despite the Council statement that in the opinion of their examiner he should be proficient after 8 weeks.

84 (Iran)

I think this work is very important. Occasionally we have had students whose English has been inadequate. This is very frustrating for staff and students; it seems to me that this problem arises from indifferent assessment before leaving their native country.

I hope that as a result of your studies more attention will be given to serious evaluation of ability in English before they join a course.

85 (Indonesia)

On this course, Mr N is being presented with many technical ideas and problems which are new to him. He reads English well enough and understands the words as words, but not the new meanings they take on in a technical context. The problems are magnified when he has to express them in spoken English and since much of the course is in seminar form, this has proved a real handicap to him. He has, however, been resolute and persevering and is slowly improving.

89 (Turkey)

The Timber Studies Course is designed for the postgraduate or experimental professional who wishes to do a specialised course. Mr O arrived with a degree in forestry and is a junior lecturer at the Black Sea

University. His chief difficulty was a complete absence of technical and scientific English which slowed his reading and required many asides in lectures. His initial lack of confidence in speaking has now worn off but out of his total time in College I estimate at least a month was wasted in practising English technical reading and probably more in adjusting to spoken English in lectures.

He is expected to undertake several visits to industrial concerns but has begged to have this delayed until March purely on account of his difficulty in following faster speakers.

92 (Tanzania)

A lazy speaker - talks through his teeth. Seems reluctant to enter into any conversation or discussion mainly because it seems too much trouble to talk English.

Question 2 - Ability to understand written English:

We feel we are not really competent to judge the above either for this student or any of the other students at Poultry School. Have however attempted a comparison between the students.

93 (Tanzania)

Very much better at understanding English than speaking it. Only now beginning to ask questions and take part in discussions.

94 (Tanzania)

Obviously trying to improve ability in English. Has shown in class that he is reading about his subject and relating what he reads to practice.

95 (Tanzania)

Will improve both speaking and writing in English. Present standard largely due to lack of confidence.

96 (Tanzania)

This student's English should be better than it is! He has been in the UK previously. It is thought that he is happy with his own standard and is unlikely to improve in the future.

1119 (Thailand)

A handicap to students is the difficulty of expressing themselves rapidly under examination conditions for theory papers. In assessing the papers no account is taken of spelling errors and peculiar phraseology.

127 (Indonesia)

This is a course for mature students holding administrative posts in the Hospital Service of their country. It involves considerable project work and discussion. Dr R's use/understanding of English is entirely adequate although the pace is slower than in a wholly English group, and he probably spends a very long time over his reading and preparation.

131 (Indonesia)

Mr K gives the impression of a potentially good student who is held back by his inability to express himself. He clearly understands a lot of what he hears but can't come back quickly with questions or comment. This is obviously unfortunate - he will probably reach pass standard in the course but will not get full benefit from discussions and seminars.

1132 (Thailand)

Dr S has an excellent command of English. She has been a visitor to this country some time ago, and has maintained her standard.

136 (Mexico)

J M came to this country with a very good command of the English language - (far in advance of other students from Mexico).

144 + 660, 661 (Algeria)

This course includes subjects such as 'verbal communication', 'Economics and Law' and 'The Professional Engineer'. The last named subject is examined by a 3 hour written paper. The course has been accepted for full exemption from Corporate Membership examinations by the IMechE and we therefore require the students to demonstrate an appropriate level of command of English at the final year examination.

My assessment of the level of English attainment of these students is therefore related to this course and not to 'conventional' engineering courses.

162 (Sudan)

Although I have marked square 2 in section 1B his ability in English is similar to the ability of many of our Scottish students.

179 (Mexico)

The assessment in relation to written English is based on the writing up of laboratory books or of other reports which he has been able to prepare at his own pace. I have no experience of his capacity to write

under examination conditions. We now have a fair amount of experience with overseas (especially Latin-American) students, and frequently find that the standard of their English may fall substantially under exam conditions so that their performance may be quite inadequate even though this would not be anticipated from the standard of their work under non-examination conditions. We believe that this may reflect their inability to take adequate notes at lectures, especially in the early part of the course, when they are meeting a lot of entirely unfamiliar scientific material and, perhaps, attempting to take notes in English for the first time, as well as the difficulty of assembling and expressing their ideas under pressure.

188 (Mexico)

My main contact with Mr L is that I will be supervising his dissertation project commencing May 75 and contact during a 13 lecture course I have given. I have not had much opportunity to assess the improvement in his English ability since he arrived, though I get the impression from others that it has considerably improved.

193 (Algeria)

Part B - more accurate to place him between paragraphs 4 and 5!

The student did not take an English assessment test as he failed to attend.

194 and 195 (Algeria)

Part B More accurate to place him between paragraphs 4 and 5!

As a result of an English assessment test, this student was advised to attend English classes. The result of the test -

	195	194
Written paper	82%	80%
Aural test	68%	70%

Advised he needs additional spelling (and vocabulary) competence.

196 (Turkey)

Mr H is on 3 months' attachment to Chemical Engineering Division Harwell prior to further training with CEGB. He undertook a 6 week course in English prior to joining us.

Dr W's judgement overleaf may be rather harsh and savours rather of intolerance.

H has been with us for less than 3 weeks. Failing giving him a comprehension test, accurate answers to the questionnaire are not possible.

213 (Thailand)

It is rather difficult to classify Mr P as between the adequate and just adequate category. His difficulty is (1) slowness in reading and writing rather than inaccuracy and (2) an indistinct pronunciation which obscures the accuracy of his speaking.

233 (Mexico)

Mr R speaks with an 'American' accent - and occasional American varieties of speech. His ability to write English is much more than adequate but not completely adequate - varying in excellence with subject and time available.

245 (Iran)

Mr N is a mature student, holding a senior position in the Ministry of Social Security in Iran. He is undertaking the 1 year diploma as a preparation for a further 1 year MSc in Social Planning for which he is very likely to be accepted. On reflection, the second alternative at B seemed a more accurate description of his abilities than A. I have noted weaknesses only in his written English, and even these are minor.

248 (Peru)

Mr A worked in the UK some years ago on secondment. He is not a typical student - he is older (in his 30s), is head of an information unit; is exceptionally fluent, but speaks in a heavily accented manner. Of the students from 'non-English' areas of the world he is probably the best I have had (ie linguistically).

253 (Nepal)

Mr S's English often contains grammatical inaccuracies but these do not impede communication. He is a fluent and lively talker and writer.

267 (Mexico)

This (and other) student has two main problems with English:

1. Reading quickly, so that he can cover the material of the course.

2. Writing essays under examination conditions, when time is again a constraint.

281 (Thailand)

Why are 'ability in English usage' and 'pronunciation' grouped together? How about those who have a sound knowledge of the language (grammar,

literature, usage of the colloquial idiom) but who, nevertheless, speak with a strong foreign accent (me, for instance)? Let us also not forget Hercule Poirot!

290 (Mexico)

Mr G's difficulties are partly ones of comprehension and partly associated with the fact that he has to translate to Spanish, think in Spanish and translate back to English. This is a problem which only time would solve - not crash courses in English. The problem for him is made more acute because he is taking a degree in which the teaching is that of course work and the outcome determined by examination, not dissertation.

1316 (Lesotho)

Her background knowledge of mathematics is not strong, but she is conscientious to make the fullest use of her opportunity to improve this, in order to tackle her individually-tailored study of the teaching of mathematics.

322 (Nepal)

The MSc course by examination involves 2 terms (Xmas and Lent), which are primarily given over to a teaching programme (lectures and practical classes). G has been very conscientious and has seen me each week about the research aspects. However these meetings have been too short for me to give well defined opinions on his ability in English vis a vis its relevance to an MSc course through the medium of English.

325 (Thailand)

Other things being equal, this student's English language ability would normally have disqualified him from admission to the course.

Obviously our evaluation of his intellectual ability, his experience/ academic references, and the value of his intended future research played a part in our assessment.

Also I have worked in the student's own country and can therefore go some way towards understanding some of the student's linguistic and cultural difficulties.

The student was made to understand and fully accepted that the qualifying exam in June 75 (what will be a linguistic test as much as a disciplinary one) may well disqualify him from continuing the course. (In which case some minor certificate may be attempted/awarded.)

These 'paralinguistic' (?) considerations seem often to be an important part of the overall assessment of a foreign student - especially from a country with a linguistic situation like that of Thailand. I need

hardly say that there is a serious problem here arising out of the fairly widespread phenomenon of English language ability being found among the 'elite' social sectors. Unfortunately 'positive discrimination' has to start at home!

332 (Japan)

Dr A brings his written work, letters, etc, to me for checking, and I am very impressed by his ability to express himself on paper. Not surprisingly, he has occasional difficulties in understanding the spoken word, but when one considers the variety of accents to be heard in this department, he really does remarkably well!

335 (Japan)

What an unsatisfactory questionnaire. We require a certain standard before admission.

348 (Iran)

A comparatively young man in comparison with some of the others - in this group. More than most he finds the amount of reading necessary to be an almost insurmountable problem.

Like the rest, the palpable improvement results from his mere presence in an English speaking community.

355 (Mexico)

1. The distinction betwen 'completely adequate' and 'adequate' escapes me. If you are adequate for some task, surely that is sufficient. 'Outstanding' might be a more appropriate word.

2. The distinction between lines 1 and 3 of Question 2 also escapes me.

359 (Colombia)

Mr C's English when he arrived to start the course in September 1974 was very difficult to understand and there was some doubt as to whether he would be capable of undertaking our Masters Programme.

Mr C had originally applied for a Diploma course in Operational Research with Public Administration. This course was cancelled at a late date by which time Mr C had left Colombia for this country. On his choice he elected to take the MBA. I feel his English would have been adequate for the more technical/mathematical course he originally applied for but he experienced considerable difficulty on our Programme which involves up to 24 hours of classroom work per week. Also there are case studies and practical role playing exercises designed to encourage students to take and defend decisions.

Mr C's ability in English inhibited him from participating to the full in some of this work but lecturers were prepared to go over the main points of difficulty with Mr C.

I think Mr C's English has improved greatly since he has been forced to discuss and argue in a class of 23 experienced business men and women all of whom have good academic qualifications.

376 (Poland)

He is a particularly persistent and intrusive specimen, so that what he lacks in linguistic ability is more than compensated by determination, and sheer attrition.

378 (Poland)

Dr F now understands spoken English very well, provided it's spoken clearly and relatively slowly. But he has considerable difficulty understanding conversation at normal speeds, and where the speaker has a strong dialect.

His spoken English is very good from the point of view of communicating ideas to the hearer, and he's seldom stuck for a word.

Written English he understands very well indeed, but he naturally finds difficulty in writing English, partly because he has little practice, but also because it's naturally more difficult.

382 (Spain)

Dr R is a little difficult to assess precisely. He is a quiet individual and his difficulties, therefore, may be more apparent than real.

393 (Turkey)

The deficiencies in Mr C's case are almost wholly in written English - largely grammatical mistakes or use of words in wrong context which do not always make clear what he is trying to express. This may prove to be a considerable handicap in written examinations when he is working within strict time limits.

1399 (Turkey)

1B. I didn't know whether you include a very slight accent as a 'minor fault' in pronunciation.

I also didn't know whether 'native speaker' refers to British as opposed to other English users, eg in Britain, but not in the USA. This girl might pass as American.

I would have thought a distinction was worth marking between usage and pronunciation.

400 (Turkey)

Q1B. Spoken English - He makes occasional errors with words with non-native pronunciations.

Written English - His grammar and vocabulary are good but he occasionally uses non-technical words in place of more appropriate technical words.

Q4. He has taken a course in English with the British Council, but it was not sufficiently advanced. He plans to take, or may currently be undertaking a more advanced class.

403 (Japan)

Mr H is an interesting case. He is a member of the Japanese diplomatic service and was educated partly in English language schools in Delhi while his father was at the Embassy there. He is therefore well above average for a Japanese: but finds academic English less familiar than diplomatic English.

501 (Argentine)

Results just satisfactory so far.

Student is aware of the problem, and is taking steps to improve his abilities in the English language.

505 (Oman)

Mr T has worked very hard but he continues to make very fundamental mistakes, particularly when writing. He certainly finds the task of learning English a difficult one and has quite a long way to go before he reaches a satisfactory level.

1508 (Bolivia)

C has a desire to communicate both professionally and socially, and this positive attitude and drive without agreession are great assets in her relationships.

509 (Brazil)

Mr de S has worked in the field with English speaking colleagues and I am convinced this is one of the most important factors. Crash courses in language are a poor substitute.

511 (Argentine)

This student would probably have improved much more rapidly in terms of English had it not been for the fact that, having taught him for 3 years in his own country, I naturally converse with him in Spanish. However, although I have marked your section 3 overlap as 'only a little improvement' this is rather unfair - but to say that his improvement has been 'considerable' would also be unfair.

517 (Ethiopia)

Our course is closely concerned with language skills, particularly in the area in which I myself lecture - editorial work. Thus any defects constitute more of a handicap, I think, than they would in many other courses. Mr F is certainly very near the box above the one marked in 1(B), and his English now appears to be improving. It seems to be the first 6 months in England which are the most difficult: last year we had a Japanese student whose progress after that time was most impressive.

522 (Brazil)

We have found Mr B's performance as a student entirely satisfactory as was the information provided by the British Council about his English ability.

525 & 526 (Jordan)

Dealing with radio subjects "technical" language problems do not arise as it is mostly symbols. Most problems occur during normal conversation, ie during coffee breaks, when many overseas visitors might sit dumbly looking on.

Both Mr Al-R and Mr A on the other hand would take and make interesting conversation although, of course, one had to be patient.

527 (Brazil)

Mr de M has brought his wife here with him and he is sensitive to her need for company, especially since she speaks little English. This retards the progress of Mr de M. In my view wives should be given English courses prior to coming to the UK (perhaps the same course as the husband) so that one can help the other with the English language.

534 (Yemen)

This student completed a HND in Building at Bristol Polytechnic before entering our course.

537 (Yemen)

Mr S is able to clarify any points arising from lectures, etc, which he may have difficulty in understanding, by discussing such points with two other postgraduate students (from Iraq) whose knowledge of English is good and who speak Arabic: these students are attending the same full-time MSc courses as Mr S.

538 (Chile)

Mr J B is not due to start his studies at The City University until 24 February 1975. He has had tuition in English elsewhere (in York and London) during the period since October 1974 (arranged by British Council).

From recent conversation with him, it appears that Mr B's ability to speak and understand spoken English has improved a little since he first came to England, but it is anticipated that he will not find it easy to follow lectures in English. Before entering the MSc Course in September 1975, he will attend a preparatory course from 24 February to end of June 1975.

539 (Chile)

This student is of above average ability and his progress in speaking and understanding English since I first met him at the beginning of the session has been remarkable. He has already presented the results of some of his experiments in a seminar and his presentation compared favourably with that of some of his English fellow students!!

I will be in a much better position to judge his written English when he presents his project report (mini-thesis) at the end of the session.

551 (Chile)

One would expect this man to speak good English since he has already taken an MSc degree in the USA in 1970 which must give him a good start over people leaving their own countries for the first time.

569 (Chile)

Candidate has made such satisfactory progress on present course that I am supporting his application to stay on for another year to study for MSc Research. I consider his command of English quite adequate for this purpose.

572 (Ethiopia)

Main problem is in understanding his verbal English - almost invariably a question from him has to be repeated, before the tutor can comprehend.

This probably embarrasses him and may well deter him at times from speaking. It is not just pronunciation but also the 'musical' mode of speaking which is strange to us.

579 (Korea)

In completing the questionnaire, I may have erred slightly towards a higher ability in English on the part of Mr K. I found myself looking for a category between those given, eg in Ques 1 part B, I have marked in the category of minor faults, probably a little generously, as he does not fit into the next lower category of many mistakes.

591 (Sudan)

Your questionnaire is far more detailed than the information normally available except in the case of 'problem students'.

596 (Ethiopia)

Would benefit greatly if prepared/able to speak out loud in a group.

Written expression far better than spoken.

598 (Chile)

Mr B's English is very good but he has had the advantage of working for several years with a team of British people in a Ministry of Overseas Development Project in Chile.

600 (Ethiopia)

T was enrolled for the one year course leading to the DMS (Agriculture). However when he commenced the course it was obvious that his ability, education and current position with the Ministry of Agriculture in Ethiopia had been overstated, and that he was unsuitable for such an advanced course.

He is now pursuing a special course at Writtle Agricultural College which is designed to suit his particular needs. He will receive some form of Certificate at the end of the academic year, but it will not be one that is nationally recognised.

602 (Ethiopia)

He appears to have no difficulty writing and understanding English, merely speaking with a foreign accent. Rather quietly spoken and does not speak up in a group.

My comment on NO IMPROVEMENT in ability since he commenced his studies

is a reflection of his ability on arrival, not any lack of progress since then.

1611 (Afghanistan)

Miss F E is a very intelligent girl, and although her English is below standard, I have confidence in her.

612 (Afghanistan)

Mr A appears to be a particularly enthusiastic individual and if his progress continues for the two year stay even remotely as well as in his first two months - then I believe he should find his final examinations within his ability.

1615 (Bolivia)

One relies on an improvement factor in English language but other factors may negative this.

I feel that English appreciation has been variable, based on personal lack of confidence at times.

1643 (Argentine)

The main problem is not that the English is inadequate in a measurable sense but that its inadequacy increases social isolation to a degree which seriously affects the happiness, the social integration and the possibility of the informal transfer of ideas. The difficulty that I have is that it is impossible to transfer ideas unless time is allowed for a careful sitting down discussion. Dr D-G has the problem of loneliness and a feeling of being out of things.

645 (Salvador)

Dr G did not begin his course here until January 6. This is a short time for evaluating progress.

He had a 3 month course in English in London before arriving here, I gather, and though he clearly made great progress during this time, he has arrived with considerable difficulties in expressing himself adequately in English despite very high intelligence.

In my opinion he is an ideal person for the course of study on which he has embarked, and thus was appropriately selected, but should have had much more English tuition before starting this course, or indeed arriving in this country.

I may add that because of his high intellectual ability I suspect his comprehension of English may soon overtake native born Britishers.

656 (Vietnam)

I cannot really combine a rating in respect of comprehension and self-expression for questions 1A and 3 - they differ so much in her case.

674 & others (Algeria)

All these students show similar abilities and their knowledge of English does not seem to hamper them in any way.

711 (Algeria)

We have given all students a lot of essay work in the first year and we are taking steps to improve command of English for all students (including the English ones!).

722 (Brazil)

Mr de S came to us after completing a course at Cambridge, his English on entry was of a good standard.

725 (Laos)

A steady improvement in his first six months of study. Initial standard quite inadequate. Likely to attain a sufficient standard by next session. This difficulty and time lag taken into account in planning structure and duration of his course. Some difficulty in finding suitable tuition, perhaps personal in origin.

727 (Mexico)

English language is not really the basic problem which this student has. He is a law student. No doubt because of his background his approach to law is almost entirely theoretical. He seems to have great difficulty in comprehending the English empirical approach. To him all that seems to matter is the view of legal writers. He has difficulty in grasping the importance which the English attach to judicial decision.

1784 (Brazil)

I fear that she will have greatest difficulty in completing written exams (1/3rd of her course assessment) as her English remains very unsatisfactory. This is odd as she is bright, has had supplementary lessons and, she says, normally speaks English with friends. My impression is that Latin Americans may have special language problems stemming from cultural strangeness. Africans, Asians, West Indians not only tend to use English as a 1st or 2nd language. They also tend to use 'clannishness' less as a response to cultural unfamiliarity.

786 (Vietnam)

My assessments may be over-generous in the sense that his English is perfectly adequate to pass the course but, I suspect, inhibits him from realising his full potential. He is a bright student who has applied to do an MSc by thesis next year. I fear that subtle problems (? language x thought) contribute to the fact that his written and verbal expression is often less logical and analytical than he is capable of.

788 (Jordan)

I feel that on 1B and 2 Mr Al-S falls in between adequate and just adequate, and the choice is rather arbitrary. He is going to take an MSc next year.

1793 (Brazil)

As Miss Dos S is not taking any examinations her ability to write English is not of any particular importance and I think that, if she takes her time, she can read English well enough for study purposes. It is her conversational ability which we find difficult, slow and uncertain. She is staying in England until the end of next term and practice may have improved her spoken English by then. The British Council is arranging further private tuition for her.

815 (Mexico)

The ability of Mexicans to understand and communicate in English seems to be considerable judging from the two students we have in Exeter. This man is marginally the better of the two. I find this disturbing, particularly in view of the proximity of an English speaking country. English assessment in Mexico appears to be totally inadequate. I wonder whether the assessors are really up to the job!

1839 (Afghanistan)

Ratings under section 2 probably err on the side of slight generosity.

840 (Venezuela)

The rating in Part 1A may be rather harsh - his ability in English is more than adequate.

845 (Solomon Islands)

Mr A's grasp of written and spoken English shows very adequate syntactic etc knowledge. Any problems reside in lack of familiarity with academic style and with organisational techniques. He has, however, consistently

held his own and in recent weeks has shown considerable improvement in his ability to organise and present his work.

1849 (Sierra Leone)

Mrs M-P is a quiet, retiring person and it was some little time before one could make any assessment of her.

Because of the cultural difference I think she found the early days here lacking in complete communication - not on the basis of language but rather that she was not sufficiently familiar with our education system and lacked the background understanding of the teaching of reading possessed by other members of the course.

She has worked hard however and is catching up and I feel that she will probably reach a sufficiently high standard by the end of the course to gain the diploma.

857 (Zambia)

Item 3 All Zambians did a special course August-September in London before university course started in October. 'A little improvement' covers both courses to date. Ability to understand spoken English and to speak English would rate 'considerable improvement' for all Zambians, combined with written ability average is 'a little improvement'.

884 (Zambia)

Mrs M's ability to write English is much affected by time available.

901 (Bahrain)

In general the British Council students who come to us are very able people and very knowledgeable in their own fields. It would be deplorable if they were to be barred by language difficulties.

There is undoubted need for much more listening to English spoken by the English rather than foreigners in their own countries. This applied to every student I have had so far. They largely waste the 1st fortnight at least getting 'on line'.

1905 (Jordan)

1. I have little experience of foreign students, so not much to compare her with.

2. To read English does require a particular feel for nuances - both in reading and writing - which many native speakers don't possess. She

has the potential but not yet, I think, enough experience...Her English would probably be entirely adequate for any other subject.

914 (Bahrain)

Mr Al-K brought his wife with him, and is living in a flat instead of in a hall of residence. This means that in a large part of his free time he is speaking Arabic with his wife whilst his fellow-students are talking English with each other.

920 (Pakistan) and 607 (Egypt)

Conversational communication is much easier than communication of a technical or scientific nature - outside of mathematics.

The course contains material which has not been met by the candidate prior to the course and which is of a technological nature. Such material has been found to pose rather severe problems, both of understanding by the student and of expressing himself.

NB The above remarks apply almost equally to Mr S (607).

921 (Morocco)

Mr S is finding the course rather difficult. This, however, is not because of any deficiency in his English but because the method of study and the amount of work required is different to anything he has experienced before.

924 (Sudan)

Mr El D's ability is more than adequate - he writes much better under pressure of time than when he has time!

1938 (Lebanon)

Mrs H showed confidence from the first day in spoken language. She reads well and shows a comprehension and understanding of English of remarkable standard.

942 (Ethiopia)

This student is exceptional in my experience in having virtually no English language problems.

948 (Brazil)

Dr A M de A has only been at this Department since the beginning of

January. It is, therefore, difficult to say how much his English has improved, but considering the short time I think there has been an improvement, chiefly one of confidence, although he could speak very good English before he came. The other improvement I have seen is that one does not now have to speak at a slightly reduced pace in order for him to understand.

I have been unable to say anything about his ability to write English since he has done no writing yet, but from his ability to read and speak I have no doubt there will be very little trouble there.

1951 (Brazil)

Comment: Tutor speaks fluent Portuguese which relieves the student of a greater incentive to speak English. This student, according to the British Council, would not have been awarded a scholarship in other circumstances (ie a special case).

1952 (Brazil)

The candidate's competent and intellectually sophisticated written work will clearly get her through the course, which is assessed by written term papers and a dissertation. If there were an oral assessment element based on active participation in seminar discussion she would be in peril, probably - though no more so than one or two of the native English speakers on the course. There has been almost no participation in the high-level cut and thrust of the seminars: how far this represents incompetence at that very exacting level of English usage (people alluding, trailing off in tentative suggestion, interrupting, often fast or impassioned), how far a difficulty of cultural adaptation from (one suspects) a more passive and didactic university style, and how far an individual or cultural (woman: South American) tendency to passivity I really can't say - attempts at making things easier seem to induce tension so I've tolerated a totally silent presence in the seminars. This week for the first time a substantive intervention was made - a speech rather than discussion-exchange, though.

957 (Chile)

The above-named student has had special difficulties with his English pronunciation and usage, since the actual musical terms used in his own country are different to those in use here. He is accustomed to the Solfege system as used in France, and has to translate this into English notations, as well as speaking another language. Being familiar with Solfege myself, I have been able to help him with this, but a teacher who did not appreciate this particular problem might find it difficult to disentangle correct answers from faulty ones. His spoken English is quite fast, but not comprehensive enough to teach in English.

959 (Chile)

His English is excellent and he can say anything he wants and enlarges his vocabulary all the time. Can follow difficult discussions.

1965 (Colombia)

Completely fluent at all levels. Her slight accent would give her away as non-English, but her intonation, pronunciation and placing of stress are always correct.

974 (France)

I think Mr M would be a very useful guinea-pig for your studies. Intelligent but with only an adequate knowledge of English. He can, however, read modern English novels fairly fluently.

975 (France)

When Dr C came here in October his knowledge of English was barely adequate, but he has attended the language laboratory daily and still does so. As a result his English has improved very considerably.

990 (Germany)

His first piece of written work showed a few small errors typical of the best German students. There was an occasional clumsiness. His second major piece of written work had no such faults.

992 (Germany)

Mr B's trouble is not that he has difficulty in handling the English language but that, having been trained in German metaphysics, he has difficulty in adapting himself to English ways of philosophical thought.

998 (Denmark)

This student is better than most native speakers. He does have a Danish accent and tiny usage errors but these are fewer than most of his English contemporaries. It would be totally misleading not to give him the maximum on all the questions. Native speakers, even graduates with 2:1 degrees, I now expect more often than not to have some language problem. I have a technical article of O's which is of first-class style, expression, and incidentally - content.

999 (Finland)

Professor S has plainly considerable grounding and experience in the area of English. While not completely fluent or idiomatic, he can express himself quite easily and well. I imagine that he has reached a level at which he would remain, in the absence of further deliberate effort at self-improvement.

2100 (Sweden)

I was a little unhappy with your own divisions and would have been happier with a 'more-than-adequate' space between 'adequate' and 'completely adequate'.

2104 (Cyprus)

Despite slight deficiencies in spoken and written English N S is a highly intelligent, intellectually enterprising man. I am impressed by the extent and depth of his reading.

2021 (Malta)

I think it's her second rather than a foreign language.

2117 (Italy)

Somewhat limited vocabulary and slow delivery, nevertheless adequate.

2028 (Spain)

Section 3. Improved in use of more colloquial English.

2152 (Japan)

By comparison with previous Japanese students whom I have taught K's command of English is improving satisfactorily. He seems to be able to understand English reasonably well - both in written and spoken form - but says that he still has to concentrate quite hard in order to keep up, and also that his command of general conversational English is rather weaker than of the vocabulary necessary for his studies.

The Japanese seem to find it difficult to learn English and by their usual standard he is doing well and certainly it is not the barrier that it is for other students I have met.

Section 6 Communicative Proficiency Measure - sample background data

Table 7.39 Places of study

Univ/College/Centre	Sample - N	Ratings - N	CPM - N
Aberdeen	12	10	8
Edinburgh	36	34	16
Moray House	25	24	18
Glasgow and Strathclyde	33	24	6
Leeds	44	39	17
Imperial College London	30	23	7
LSE	18	15	1
Manchester and Salford	71	55	21
Southampton	12	3	1
	281	227	95

Table 7.40 Distribution by geographic area

	Areas	CPM Freq	CPM Freq %	Whole sample Freq %
1.	Africa (South of the Sahara)	12	12.6	11.0
2.	Middle East & North Africa (including Sudan)	18	18.9	24.4
3.	South, Southeast and East Asia	28	29.5	22.6
4.	Latin and Central America	31	32.6	27.9
5.	North and South West Europe	2	2.1	6.6
6.	East and Southeast Europe	4	4.2	7.5
		95	100.0	100.0

Table 7.41 Distribution by subject area

	Subject Area	Freq	Freq %	Whole sample Freq %
1.	Agriculture, etc	4	14.2	8.4
2.	Arts	-	-	2.9
3.	Medicine, etc	5	5.3	7.3
4.	Pure Sciences	9	9.5	11.4
5.	Education	33	34.7	16.2
6.	Engineering, etc	9	9.5	21.8
7.	English Studies, etc	3	3.2	6.2
8.	Social Sciences (prof)	30	31.6	18.0
9.	Social Sciences (acad)	2	2.1	7.3
10.	Miscellaneous	-	-	0.5
		95	100.0	100.0

Table 7.42 Distribution by level of study

	Subject Area	Freq	Freq %	Whole sample Freq %
2.	Pre-university	-	-	1.2
3.	First degree	-	-	5.5
4.	Masters by tuition	16	16.8	23.3
5.	Research (degree)	5	5.3	16.0
6.	Professional diploma	74	77.9	41.6
7.	Academic attachment	-	-	9.5
8.	Professional attachment	-	-	2.9
		95	100.0	100.0

Table 7.43 Distribution of EPTB and BCSA results

	EPTB		Freq	Freq %	Whole sample Freq %
1.	less than 34.0	English inadequate	5	13.5	16.7
2.	34.0 to 39.9	English tuition needed	27	73.0	50.8
3.	40.0 to 45.9	English adequate	5	13.5	26.0
4.	46.0 and above	Unquestionably adequate	-	-	6.4
		Total	37	100.0	100.0
		Mean	37.62		38.32
	BCSA				
1.	C+ and below	English probably inadequate	9	15.5	21.6
2.	B, B+ (low)	English tuition needed	12	20.7	25.6
3.	B+ (high), A	English adequate	37	63.8	52.8
		Total	58	100.0	100.0
		Mean	low B+ (3.43)		low B+ (3.34)

Table 7.44 Distribution of oral and writing assessments

Oral Grade		Frequency	%	Whole sample Frequency %
D+		1	1.4	1.9
C		2	2.9	8.6
C+		3	4.3	5.3
B		26	37.1	34.7
B+		5	7.2	9.8
A		33	47.1	39.7
	Total	70	100.0	100.0
	Missing	25	Mean B+	Mean B+
Writing Grade				
D+/D		-	-	2.2
C		11	17.7	20.2
C+		3	4.8	9.2
B		29	46.8	35.5
B+		2	3.2	3.8
A		17	27.4	29.1
	Total	62	100.0	100.0
	Missing	33	Mean B	Mean B

Table 7.45 Length of remedial English tuition

No of weeks	Frequency	%	Whole sample Frequency %
Up to 4	17	24.3	30.4
5 to 8	25	35.7	31.1
9 to 12	27	38.6	32.8
13 or more	1	1.4	5.7
Total	70	100.0	100.0
None, as % of sample	25	26.3	36.9

Section 7 Communicative Proficiency Measure - results

Table 7.46 Summary of results of the cloze tests

	Test AC30	CC30	DC25	MC25	TC
N	86	88	87	79	94
Mode	12	13	10	8	27
Mean	11.4	14.1	9.6	11.1	42.0
(Mean %)	(38%)	(47%)	(38.4%)	(46.2%)	(38.5%)
SD	4.7	4.5	3.7	5.3	15.2
Minimum score	1	4	1	1	14
Maximum score	23	30	25	23	85
Reliability	.68	.64	.58	.79	.89

Notes: N = number of students taking the test or subtest.
Test codes indicate the number of items in each cloze test except for MC25 which contained only 24 items.
Reliability was worked out according to the Kudor Richardson 21 formula. Workings appear in Table 7.49.1
TC = total cloze test scores.

Table 7.49 Pearson correlation coefficients for the cloze tests (CPM)

	AC	CC	DC	MC	TC
AC	1.0000 (0) P=*****	0.5685 (81) P=0.000	0.4544 (79) P=0.000	0.5587 (72) P=0.000	0.7759 (86) P=0.000
CC	0.5685 (81) P=0.000	1.0000 (0) P=*****	0.4574 (81) P=0.000	0.3643 (74) P=0.001	0.6618 (88) P=0.000
DC	0.4544 (79) P=0.000	0.4574 (81) P=0.000	1.0000 (0) P=*****	0.3369 (75) P=0.002	0.6167 87) P=0.000
MC	0.5587 (72) P=0.000	0.3643 (74) P=0.001	0.3369 (75) P=0.002	1.0000 (0) p=*****	0.7712 (79) P=0.000
TC	0.7759 (86) P=0.000	0.6618 (88) P=0.000	0.6167 (87) P=0.000	0.7712 (79) P=0.000	1.0000 (0) P=*****

Table 7.49.1 Workings of reliability coefficients for cloze tests

KR21 $$Rel = 1 - \frac{M(n - M)}{ns^2}$$

Where M = mean
n = number of items
s = standard deviation

AC 30 $$Rel = 1 - \frac{11.4(30 - 11.4)}{30 \times 4.7^2} = 1 - \frac{11.4 \times 18.6}{30 \times 22.09}$$

$$= 1 - \frac{212.04}{662.7} = 1 - .32$$

$$= .68$$

CC 30 $$Rel = 1 - \frac{14.1(30 - 14.1)}{30 \times 4.5^2} = 1 - \frac{14.1 \times 15.9}{30 \times 20.25}$$

$$= 1 - \frac{224.19}{607.5} = 1 - .36$$

$$= .64$$

DC 25 $$Rel = 1 - \frac{9.6(25 - 9.6)}{25.x\ 3.72^2} = 1 - \frac{9.6 \times 15.4}{25 \times 13.63}$$

$$= 1 - \frac{141.8}{340.75} \quad 1 - .42$$

$$= .58$$

MC 25 $$Rel = 1 - \frac{11.1(24 - 11.1)}{24 \times 5.3^2} = 1 - \frac{11.1 \times 12.9}{24 \times 28.1}$$

$$= 1 - \frac{143.19}{674.4} = 1 - .21$$

$$= .79$$

TC (Total) $$Rel = 1 - \frac{42(109 - 42)}{109 \times 15.2^2} = 1 - \frac{42 \times 67}{109 \times 231}$$

$$= 1 - \frac{2814}{25179} = 1 - .11$$

$$= .89$$

Table 7.50 Distribution of levels for writing

	Level	Frequency	%
1.	No communication	1	1.1
2.	Inadequate	4	4.4
3.	Serious weaknesses	25	27.8
4.	Some weakness but adequate	22	24.4
5.	Few weaknesses, very competent	23	25.6
6.	Close to N-S proficiency	15	16.7
	Total	90	100.0
	Missing	5	

Table 7.51 Distribution of levels for oral

	Level	Frequency	%
1.	No communication	1	1.2
2.	Inadequate	9	10.6
3.	Serious weaknesses	13	15.3
4.	Some weaknesses but adequate	25	29.4
5.	Few weaknesses, very competent	22	25.9
6.	Close to N-S proficiency	15	17.6
	Total	85	100.0
	Missing	10	

Table 7.52 Inter-test correlations - CPM

	C	W	S	WS
C	-	.506	.488	.551
W		-	.746	.928
S				.94

N varies from 80 to 90. All coefficients are highly significant

C denotes total cloze score, W the writing test level, S the interview level, and WS the combined writing and interview level.

Table 7.53 Test results recoded to a common scale

Scale 1 - inadequate proficiency
2 - only just adequate
3 - adequate
4 - completely adequate

Recoded results

Subject	Total cloze	Writing	Oral	Overall
509	3	1	1	1
11	3	2	1	(2)
25	3	3	3	3
48	3	2	3	3
56	3	3	3	3
64	2	2	1	2
81	3	3	2	3
121	3	3	4	3
140	2	2	2	2
181	4	2	3	(3)
184	3	3	4	3
232	2	2	3	2
233	3	4	4	4
236	2	1	2	2
237	1	2	1	1
239	2	3	-	-
241	2	-	3	-
253	3	3	-	3
1272	4	4	4	4
286	3	3	3	3
288	3	3	3	3
1287	3	3	3	3
297	3	3	3	3
298	2	2	1	2
299	1	2	3	(2)
300	1	1	1	1
301	2	3	3	3
302	3	3	3	3
303	3	2	3	3
304	2	2	2	2
305	1	2	3	(2)
312	3	3	3	3
1313	3	3	3	3
314	3	3	3	3
1315	2	3	3	3
320	3	4	4	4
327	3	3	3	3
1347	2	2	2	2
519	3	-	3	3
579	2	2	2	2
583	4	3	3	3
584	1	3	2	(2)
590	3	4	4	4
1593	1	-	3	-

Subject	Total cloze	Writing	Oral	Overall
1601	3	3	2	3
606	4	4	4	4
610	–	–	3	–
650	3	2	2	2
701	3	3	1	3
725	2	2	2	2
727	3	3	3	3
1754	4	4	4	4
1755	3	3	3	3
777	2	3	3	3
778	4	3	3	3
781	2	2	3	2
1782	1	–	1	1
782	3	3	4	4
783	2	2	2	2
1794	3	3	4	3
1799	2	1	–	–
800	2	3	–	–
1803	2	2	3	2
802	3	3	3	3
811	3	2	2	2
812	3	1	2	(2)
813	1	1	3	1
822	3	3	3	3
823	4	3	3	3
1838	3	4	–	–
1839	1	2	3	(2)
840	3	4	3	3
1841	3	3	3	3
842	3	2	3	3
1843	2	3	3	3
844	2	2	2	2
845	3	3	–	3
846	3	4	–	–
847	3	4	–	–
848	3	4	3	3
1880	1	3	3	3
886	3	3	3	3
892	2	4	–	–
2161	1	3	3	3
2162	2	3	3	3
924	3	3	3	3
927	1	3	3	3
1941	4	4	4	4
945	3	3	4	3
1961	4	4	4	4
963	3	4	4	4
975	2	2	1	2
1989	4	3	4	4
2130	1	3	–	–
2139	4	3	3	3

Table 7.54 Distributions of categories of adequacy - summary of Table 7.53

	Reading	Writing	Speaking
1. Inadequate	13 (13.8%)	5 (5.5%)	10 (11.8%)
2. Only just adequate	24 (25.5%)	25 (27.8%)	13 (15.3%)
3. Adequate	46 (49.0%)	45 (50.0%)	47 (55.3%)
4. Completely adequate	11 (11.7%)	15 (16.7%)	15 (17.6%)
Total	94 (100.0%)	90 (100.0%)	85 (100.0%)
Missing	1	5	10

Section 8 Selection of essays by level

Level 6	No 1961	-	female, following education course in Edinburgh; age 25; from Mexico
	No 320	-	male, following Masters in linguistics in Leeds; age 39; from Bangladesh
Level 5	No 845	-	male, following a course at the Institute of Education,Leeds; age 24; from Solomon Islands
	No 1315	-	female, following education course in Edinburgh; age 38; from Botswana
Level 4	No 25	-	male; following a course in physical sciences in Aberdeen; age 32; from Indonesia
	No 584	-	male; following a course in agriculture; age 37; from Cyprus
Level 3	No 842	-	male; following a course in public administration in Manchester; age 29; from Mexico
	No 48	-	male; following a course in engineering in Glasgow; age 28; from Sudan
Level 2	No 300	-	male; following a course in administration in Manchester; age 27; from Peru
	No 1799	-	female; following a course in education in Leeds; age 24; from Bahrein
Level 1	No 812	-	male; following a course in statistics in Aberdeen; age 27; from Jordan

Photocopies of the essays are presented in the above order on the following pages.

1961

Life in Britain has been quite difficult for me. Before I left Mexico I thought things were going to be very different.

I have been abroad several times during my life so I had was sure this was going to be another magnificant and intresting experience. But things began going wrong as soon as I arrived in Edinburgh. For the first time in my life I could not adapt myself to a new and situation. I just could not understand myself because I had never felt this way before. Fortunately the first term went by very quickly and when the holidays arrived I was sure things were going to be better. I went to Switzerland where I had a great time and tried to forget how lonely and miserable I had been feeling before. At the end of the holiday I told myself I would try to be more cheerful during the second term and concentate on the course so that life would not be so difficult.

But when I came back in January I felt worse and worse every day. I had never been so nervous before. I tried to find the reason why I was feeling that way but I couldn't. Suddenly on the 12th of February my mother phoned me from Mexico to tell me that my father had died the night before. I could hardly believe her. I did not

know what to think or how to react. I decided I had to go home for a few days, I just could not stay here. I went to Mexico, saw my family and ten days later ~~I~~ I came back. Of course I did not feel like coming back at all but I had to in order to finish the course. Since then I have been feeling very, very lonely and depressed; a lot more than before.

Perhaps if my father had not died, things would have been different. I would have tried to see the nice part of this experience but things unfortunately I have been quite pessimistic. I just lack the strength to try to be cheerful. I have the impression that during this year I have not been myself. Nevertheless it has been a very good experience and anyway that is what life is like. There are good and bad moments and we have to accept whatever happens.

The death of my father has been such a shock for me that when I think of my stay in Edinburgh it is all I can think of. The course, the people, the city, everything else comes next.

Before I came to Scotland, I thought I was going to learn a lot about such things as linguistics and phonetics but now I have realized I have learned a lot more about life itself.

32o

From the beginning I have been very much interested in the education system at the university and in the secondary schools. For I believe that it is the system which is more interesting and significant than anything else. The education system or a part of it reveals a great deal about the political and social systems of a given country because all education systems are products or extensions of the political and social systems.

What I find rather disturbing is the attitude of both teachers and students in the university. Generalisations are misleading but we find the teachers authoritarian and this makes the students rely heavily on memorization of the lectures notes and books. The students seem to be more concerned with what the lecturers think about a problem than with finding, or trying to find, an answer of their own produced through the process of inquiring, independent thinking and reflection. This is a sad state of affairs at any level of education, especially at the university which is a place for open and free thought, intellectual and aesthetic growth and search for truths.

The question then arises? why is it so? why are the lecturers so authoritarian and domineering? why are the courses designed in such a way which gives

the students little opportunity to think independently and grow intellectually? Why should it be like this in a democratic society? These are complex questions involving many social, cultural and historical factors. However, one can identify and notice a pervasive streak of conservatism and authoritarianism in the English way of life, especially in the education sector of this country.

I have been interested in this aspect of education here partly because by profession I am a teacher. But I also believe that one should try to learn more about systems and values and attitudes underlying them. Another thing I have noticed is what may be called conceits in many teachers — an unshakable belief that the quality of education in this country is better than that in some other advanced countries. I should point out, however, that there are many exceptions, which is only natural.

845

C. SOLOMON ISLANDS

INSTITUTE OF EDUCATION (CEU)
LEEDS UNIVERSITY

LIFE IN BRITAIN:

Living in a strange country I think depends on varies from time to time ~~according~~ depending on what kind of mood a person is in. Take my case for instance, there are times during my stay in Britain so far that I feel Britain is a terrible place. cold, people are very reserved and so hostile and there are times when I am in the swing and forget al about home and enjoying myself more than I would have in a home situation. it is therefore not so much a question of what life in a new situation is like but how one treats life in that situation.

When I arrived in Britain last Summer, I was so fascinated by the huge buildings, beautiful parks, Cinemas, Theatres and the gay lights of London that not at all did I felt like thinking back of my own Country. there was not much really for me to worry about during my first five months in London. My pocket money was given weekly, my food cooked for me, my bed done for me and felt I was bit of an important man or Something of that sort.

When September finally arrived and I was moved over to Leeds, my new residence shewin and had to get down to Some Solid work, my attitudes of Britain changed. In Leeds, I had to worry about, shopping, which I am not good at, Cooking for my self, Washing up, doing my own room, paying of my rents. Cost of things were soaring high and in addition to all that, the Cold weather added to my frustrations. Mail from home become irregular and my idea of Britain changed. But on the whole life in Britain is full of fun, entertainments, lots of things to see. but it takes time to assimilate oneself in it.

1315 Botswana.

My ... interesting ... Experience

It is now six months since I've been doing my PHT Course at Moray House college of Education. My course is quite intresting but there nothing I enjoyed most like Schools Observation.

It was on the 7th and 8th weeks of term the fist term when my group visited Schools. I went to B. Haileybend Primary School. When I got there I was highly impressed by the buildings, the classrooms are ~~is~~ spacious. In fact there are no ~~fot~~ partitions between the classroom. I was told such are called "Open Plan Classrooms". The classrooms allow more 'learning corners' to be set up within the room as well as display of pictures, other teaching aids and cupboards or shelves for class libraries. Above all children are free to move around in the classroom.

I was also struck by the class organisation in almost all the classes I ~~was~~ visited. The mere arrangement of tables. Children sit in groups, work together unlike the old system ~~from~~ I am used to of arranging tables or desks in rows facing the teacher. I soon found out that here children play an active part in the part of their education. The teacher guides and helps not spoon feed. Children feel responsible at a very early age.

At the time I have been thinking That everything is provided for teachers on the list of teaching aids. but I discovered that I was wrong; Teachers make their own teaching aids. Children too are enc

to do their which are later. displayed.
This encourage children to do their work.
Finally I hope what I have seen I'll
try to introduce to my own country,
At the same time I feel the Scottish Schools are
rather well equipped.

25

My life in Britain.

I am one among those who are lucky to have the chance to stay in this country. Here, I can see as well as experience many new things which I only used to hear before. Certainly there some disadvantages yet the advantages are still a lot more.

If I mentioned about disadvantages I meant that being a married man it is not easy to be away from my family. Also, being a man who was born in a tropical country, I sometimes find that the weather is a little bit annoying, especially in winter. Such problems, are however, very personal and not so important and I think I will be able to overcome advantages, as I have mentioned above, as the

The most important thing, I can see how people in this country lead their ~~L~~ lives. I found that most of them are very helpfull indeed. Also the way they run the country is something that is worth learning. They have made life so easy and very convenient, something which is very hard to find in my country.

Being used to such an easy life in this country, I think I will ~~fe~~ a little frustrated when I get back home to my country. I have to face it anyway and I always try not to think about it. The best way to do is to learn more and more which may be useful to my next life in my home country as well to the country where I live in.

Economy of Cyprus:

Cyprus is situated in the east of the Mediterranean Sea, and has a typical winter rainfal and dry summer climate. The winters are rainy and mild cold, but summers are dry and hot.

Population of Cyprus is about 600.000 of people

Economy of Cyprus ~~depends~~ mainly depends on agriculture. Among agriculture products ~~are~~ citrus fruits orange, grapefruit lemons, onions, grapes are mainly exported. water melons, sugar melons and potatoes are also produced and potatoes are exported also.

Animal production in Cyprus is not a source of income which attribute to economy of Cyprus. But is self efficient for its meat ~~production~~ requirements. cattle population is about 60.000 and sheep and goat population is about 600.000. ~~Among the animal~~ cattle Dairy cattle has ~~been~~ increased up to 20.000 which these cattle were imported from Holland and Britain, mainly they are Frizians type

Light industry has been improving ~~is~~ for 15 years. Among light industry plastic products are important and is exporting to middle east countries. Another source of income among the industrial products is wine.

Tourism in Cyprus has been developing for 10 years. Cyprus has got sunshine all the yearrounds. Nice beaches and sunshine is attracting thousand of Tourist every year. Tourists enjoys themselves in sunshine, and nice beaches.

Cyprus is also a historical country. Ancient monuments museums are very attractive.

~~in winter snow fall~~ Tourists also enjoys themselves in mountains where in winter time there are snow. Holland ~~skii~~ is and Tourists likes to skiying on snow.

Natural sources of Cyprus includes iron pyrites and copper and also asbestos. These products are also exported and are considered as sources of income.

Among the animal productions milk, meat and eggs are sufficient to Cyprus ~~and milk is~~ drunk milk is drunk as pasteurized milk and also is used in cheese making. But cheese is not enough for and is imported. Because number of Dairy cattle is not self sufficient.

842

One of the most important aspects of my life in Britain has been, in fact, the experience of living for the first time in my life by my own. While I was here alone, I had to fight with myself in order to overcome some common obstacles such as: the difficulty to communicate with native people, and to find new friends as well. — Other important aspects of my Stayng here, is included the challenge to succeed ~~that~~ in my Studies. ~~or the~~ ~~same~~ So as to Succeed in my Studies I had to deal not only to my Subjects but also as a human being who lives surrounded by other human beings & had to find new friends in order to interchange ideas, ~~ideas~~ attitudes this experience has helped me to understand better myself, also has allowed me to know better my own country - because being out of problems ~~classify~~ sometimes Gets someone more objective and ~~a~~ critic.

48

Every one who goes to a new country will
mos probably put in most certain things that he
is going to expect there. In fact I spent in London
only four nights and that was on my first
arrival; and i didnot see Englan ever since.
In Scotland really I begin to understand
somethig about the people of scotland totally
different from what I was first impressed with.
And this discovery is mainly due to my
interest in knowing some of the scottish
tradtions.

What I see here in glasgow will never
give me the real scottish person. So Unfortunately
only throug the Television that I found the
real scottish man. Their songs are totally differd
from what we see in the dance halls. i tried
to read for a poet (Albert Burns) but in Jail
there is a difficully in understanding because
most of his words are scottish! It has a
name but I do not remember.

Another thing that also interests me
is the beautifil Country side. Really it is
very fine. I visited Edinburgh and Dundee
and iam really pleased.

I think it worth while mentioning that,
despite the strikes from time to time, every thing is
going in order. In fact I must be clear and
say that I mean the way in which the office
work is conducted. During the working hours
every person perform his work with honesty.
in the developing Countries such problems arise
and they are really a challenge. The remedy however

will not be easily and it needs time.

Finally, although I tried to express what I feel here, but I would like to mention that I think myself still new here and this is because I spent most of the time since from my arrival since then I is following a tough course. Now I am free to get on with my project and I will be having enough time to see more.

300

One of the most important xpereice in my life was the day that I arrived to London. I was complety lost, and I didn't speak English at all the few words ~~I~~ which I knew wasn't enough. for xspres myself. However always you find somebody who help you during my studies in Britain I found a lot of very kindly people.

Another important aspect of my xperience in Britain was the way of life and the way how culd I adapted to it. As you know the costums changes among the countries, and many things which you use in your own country are ~~not~~ xen it. in different form.

1799

Thinking of my work, I can ~~see~~ see that there are so little of interests. The most thing I could say was interesting is my our visits to school. As a part of my course, we had to visit different type of school, the ~~is~~ primary, middle, and secondary, ~~but this would last, I hope it~~ Unfortunatly, we stopped of going to these schools nowadays. Why - I don't know; we have choice to ask our tutors to go to some school if we wish to. but the problem is how ~~to~~ I can I pay, ~~that~~ I want to go to. ~~school~~ a particular school? ~~but~~ I don't know that school, I hoped to go one day to be a primary school and teach there, but unfortunatly I can't.

What ~~make~~ me interested, is the life of school the relation between the teacher and pupils. The way of teaching them, even sometimes, teachers used old methods of teaching especially in teaching a foreign languages. but I found that ~~by~~ the children have so much freedom which may spoil them. You can't find this in my country, there is a big difference between the schools here and there. That is what make ~~me find out the difference~~ it interesting for me, and lead me to find out the difference between both of them:

The other thing which make it to me interesting is the relationships between the teacher and pupil as I mentioned before. this ~~make~~ made me ~~for~~ interested to write something about it and compare.

Actually, writing a report for my self about this ~~was~~ interesting ~~thing~~ of school life ~~made~~ was so beautiful, it really encourages me to take these reports to my country next year and work on it and do ~~the~~ similar thing in my schools if I could

(40)

812

Two things which I saw them very interested: the way of working during the week, daily from 9.0 am to 5 P.M. and this is very nice to maintain the progress which GREAT BRITIAN acheived. and we were in need of it, to develop our countries. And the work means the work every thing has it is own limits. And in the work time we should work. This is valuable to me so as to try to do my work as you do here and as it should be done.

(10 minutes)

Section 9 Content of two interviews

Extensive notes were taken during some interviews. Information obtained during two such interviews is presented below. The left-hand column indicates the topics discussed, the centre column contains the information given by a Japanese male student studying in Leeds, and the righthand column the information given by a Chilean male student at the Imperial College of Science, London.

Topic	Japanese student	Chilean student
1. Home town	Nagoya: between Tokyo and Kyoto. There is a fast train that travels at 250 kph. Noise pollution, but comfortable inside the train.	Santiago, temperate climate but warmer than Britain. Lowest temp is -5°.
Family	2 children. Coming to Britain in the summer.	1 daughter. Wife speaks no English.
Mail	Post takes 5 to 7 days.	
2. Accommodation	Devonshire Hall of Residence. 3 meals per day, 4 formal dinners per week. Most other residents are freshmen or Asian p-g's. Has a normal bedsitter, furnished. Had requested such accommodation. Occasional contact with British students.	Living in flat in Clapham Common. Wife copes with the shopping. Would recommend scholarship to single people only or have married scholarships. Unhappy at leaving family at home.
3. What has struck you most about UK?	Old things. Buildings. Not renewed as in Japan. Wood houses in Japan compared with stone houses in UK. Change comes more slowly in UK.	How nice, quiet, slow the pace of London is. The people are polite and helpful.
4. Course. Describe work, project.	TEO Diploma. Theory and practical.	MSc Chem Engineering. 8 lecture courses in six months. Qualifying exam in April. Then thesis on automatic control. Chemical Engineering deals with chemical processes to obtain products. Control necessary for eg heat. Desirability of avoiding manual control.

5. Will this course prove useful on return to home country?	Has changed basis of teaching English. Many obstacles to changing teaching methods in Japan. Feels more adequate to use English in class. Important exams now in Japan - pre-senior high pre-university	Yes. Hopes to become university lecturer. Many new techniques learned. Approach very different in Chile.
6. English test before departure? English language courses on arrival? Comments? Problems?	Test in Tokyo. Listened to tapes. Had course in London. So poor that he wondered if it was just a summer job! The U of Leeds course was very useful.	3 months course at S Devon Tech. Too many disparate levels. Found listening to radio and TV a problem. Was told it will take time to improve!
7. Subjects thanked by interviewer.		
Awarded	Level 5	Level 2

APPENDIX V SECOND INVESTIGATION: FURTHER ANALYSIS OF RESULTS

Section 1 Tutors' ratings and CPM

Table 8.1 Correlation coefficients: tutors' ratings with CPM

		C	W	S	WS	O
Tutors' Ratings	R1A	0.4107 (89) P=0.000	0.5433 (86) P=0.000	0.5308 (80) P=0.000	0.5697 (76) P=0.000	0.5207 (79) P=0.000
	R1B	0.4883 (88) P=0.000	0.5981 (85) P=0.000	0.6058 (79) P=0.000	0.6238 (75) P=0.000	0.6124 (78) P=0.000
	RT	0.5114 (82) P=0.000	0.6561 (79) P=0.000	0.6104 (73) P=0.000	0.6559 (69) P=0.000	0.5765 (72) P=0.000

Key:
- R1A = Rating, question 1A
- R1B = Rating, question 1B
- RT = Total skills rating (question 2)
- C = Total cloze test score
- W = Writing level
- S = Interview level
- WS = Composite Writing/Interview level
- O = Overall CPM level

Table 8.2 Distribution of adequacy according to tutors and CPM

Tutors' ratings	CPM overall categories* 1	2	3	4	Total	%
a. Identical with CPM	0	7	30	6	43	53.8
b. One category higher than CPM	3	10	9	-	22	27.5
c. Two categories higher than CPM	2	-	-	-	2	2.5
d. One category lower than CPM	-	4	4	4	12	15.0
e. Two categories lower than CPM	-	-	1	-	1	1.2
Total	5	21	44	10	80	
%	6.2	26.2	55.0	12.5		100.0

* (1 = inadequate, 2 = only just adequate, 3 = adequate, 4 = completely adequate)

Table 8.3 Contingency tables: tutors' ratings and CPM results

Tutors	CPM 1	2
1.	0*	5
2.	5	70*

* confirmed expectancies rate = 87.5%

Section 2 Pre-departure assessments and tutors' ratings

Table 8.4 Correlations: pre-departure assessments with tutor' ratings

		EPTB Pt 1	BCSA total
Tutors' Ratings	R1A	0.3058 (311) P=0.000	0.3643 (368) P=0.000
	R1B	0.2703 (311) P=0.000	0.3435 (371) P=0.000
	RT	0.3568 (291) P=0.000	0.3592 (352) P=0.000

NB This is an alternative presentation to that set out in paragraph 8.2.1

Table 8.5.1 Correlations: EPTB sub-tests with tutors' ratings

		RL	RS	RR	RW	RT
EPTB	D1	0.0633 (281) P=0.145	0.0770 (281) P=0.099	0.1402 (272) P=0.010	0.0996 (264) P=0.053	0.1146 (263) P=0.032
	D2	0.1675 (281) P=0.002	0.1591 (281) P=0.004	0.2297 (272) P=0.000	0.2315 (264) P=0.000	0.2258 (263) P=0.000
	D3	0.2172 (286) P=0.000	0.2549 (286) P=0.000	0.2426 (277) P=0.000	0.2882 (269) P=0.000	0.2986 (268) P=0.000
	D4	0.2130 (286) P=0.000	0.3254 (286) P=0.000	0.2337 (277) P=0.000	0.3432 (269) P=0.000	0.3308 (268) P=0.000
	DT	0.2587 (311) P=0.000	0.3335 (311) P=0.000	0.2936 (301) P=0.000	0.3433 (292) P=0.000	0.3568 (291) P=0.000
	D5	0.2635 (130) P=0.001	0.3451 (130) P=0.000	0.3805 (129) P=0.000	0.4023 (127) P=0.000	0.4021 (127) P=0.000
	DO	0.3109 (169) P=0.000	0.3614 (169) P=0.000	0.2880 (161) P=0.000	0.2988 (158) P=0.000	0.3605 (157) P=0.000
	DW	0.2398 (151) P=0.002	0.3394 (151) P=0.000	0.2532 (144) P=0.001	0.3325 (140) P=0.000	0.3496 (140) P=0.000

Key: D1 = Phonemic discrimination
D2 = Intonation
D3 = Reading Comprehension
D4 = Grammar
DT = EPTB Pt 1 Total
D5 = EPTB Pt 2 Reading Speed
DO = Oral
DW = Writing

Tutors' Ratings:
RL = Listening
RS = Speaking
RR = Reading
RW = Writing
RT = Total skills

Table 8.6.1 Correlations: BCSA sub-tests with tutors' ratings

		RL	RS	RR	RW	RT
BCSA	D1	0.2846 (366) P=0.000	0.3129 (367) P=0.000	0.1389 (360) P=0.004	0.2579 (348) P=0.000	0.2907 (347) P=0.000
	D2	0.2439 (366) P=0.000	0.3327 (367) P=0.000	0.1654 (360) P=0.001	0.2323 (348) P=0.000	0.2871 (347) P=0.000
	D3	0.2164 (349) P=0.000	0.2494 (350) P=0.000	0.2320 (343) P=0.000	0.2488 (332) P=0.000	0.2716 (331) P=0.000
	D4	0.2835 (349) P=0.000	0.3353 (350) P=0.000	0.2163 (343) P=0.000	0.3437 (332) P=0.000	0.3572 (331) P=0.000
	DT	0.3043 (370) P=0.000	0.3700 (371) P=0.000	0.2237 (364) P=0.000	0.3228 (353) P=0.000	0.3592 (352) P=0.000

Key: D1 = Listening
D2 = Speaking
D3 = Reading
D4 = Writing
DT = Total Grades

Tutors' Ratings:
RL = Listening
RS = Speaking
RR = Reading
RW = Writing
RT = Total Skills

Table 8.7.1 Cross-tabulation: EPTB (GRDT) by tutors' ratings (R1A)

Count / Row PCT / Col PCT / Tot PCT		R1A 1.0 Inadequate English	2.0 Only just adequate	3.0 Adequate	4.0 Completely Adequate	Row Total
GRDT 33.9 and below	1.	3 5.8 33.3 1.0	20 38.5 24.4 6.4	21 40.4 12.7 6.8	8 15.4 14.8 2.6	52 16.7
34.0 to 39.9	2.	5 3.2 55.6 1.6	50 31.6 61.0 16.1	87 55.1 52.4 28.0	16 10.1 29.6 5.1	158 50.8
40.0 to 45.9	3.	1 1.2 11.1 0.3	9 11.1 11.0 2.9	52 64.2 31.3 16.7	19 23.5 35.2 6.1	81 26.0
46.0 and above	4.	0 0.0 0.0 0.0	3 15.0 3.7 1.0	6 30.0 3.6 1.9	11 55.0 20.4 3.5	20 6.4
Column Total		9 2.9	82 26.4	166 53.4	54 17.4	311 100.0

Chi square = 44.04387 with 9 degrees of freedom
Significance = 0.0000

Table 8.7.2 Cross-tabulation: BCSA (GRST) by tutors' ratings (R1A)

Count / Row PCT / Col PCT / Tot PCT		R1A: Inadequate English 1.0	Only just adequate 2.0	Adequate 3.0	Completely Adequate 4.0	Row Total
GRST 11.3 and below	1.	3	27	44	6	80
		3.7	33.7	55.0	7.5	21.7
		42.9	37.0	24.4	5.6	
		0.8	7.3	12.0	1.6	
11.4 to 13.9	2.	3	25	47	19	94
		3.2	26.6	50.0	20.2	25.5
		42.9	34.2	26.1	17.6	
		0.8	6.8	12.8	5.2	
14.0 and above	3.	1	21	89	83	194
		0.5	10.8	45.9	42.8	52.7
		14.3	28.8	49.4	76.9	
		0.3	5.7	24.2	22.6	
Column Total		7	73	180	108	368
		1.9	19.8	48.9	29.3	100.0

Chi square = 50.74472 with 6 degrees of freedom
Significance = 0.0000

Table 8.7.3 Cross-tabulation: EPTB (GRDT) by tutors' ratings (R1B)

Count / Row PCT / Col PCT / Tot PCT	R2 (R1B) 1. More proficiency essential 2.0	2. More proficiency desirable 3.0	3. Minor deficiencies 4.0	4. Adequate 5.0	5. Near native speaker 6.0	Row Total
GRDT 1. 33.9 and below	5	18	9	19	1	52
	9.6	34.6	17.3	36.5	1.9	16.7
	20.0	26.1	12.5	14.6	6.7	
	1.6	5.8	2.9	6.1	0.3	
2. 34.0 to 39.9	16	40	41	59	2	158
	10.1	25.3	25.9	37.3	1.3	50.8
	64.0	58.0	56.9	45.4	13.3	
	5.1	12.9	13.2	19.0	0.6	
3. 40.0 to 45.9	3	9	20	38	11	81
	3.7	11.1	24.7	46.9	13.6	26.0
	12.0	13.0	27.8	29.2	73.3	
	1.0	2.9	6.4	12.2	3.5	
4. 46.0 and above	1	2	2	14	1	20
	5.0	10.0	10.0	70.0	5.0	6.4
	4.0	2.9	2.8	10.8	6.7	
	0.3	0.6	0.6	4.5	0.3	
Column Total	25	69	72	130	15	311
	8.0	22.2	23.2	41.8	4.8	100.0

Chi square = 39.49419 with 12 degrees of freedom
Significance = 0.0001

Table 8.7.4 Cross-tabulation: BCSA (GRST) by tutors' ratings (R1B)

Count R2 (R1B)

	Row PCT / Col PCT / Tot PCT	Totally inade-quate 1.0	More profy essen-tial 2.0	More profy desir-able 3.0	Minor defic-iencies 4.0	Ade-quate 5.0	Near native speaker 6.0	Row Total
GRST 11.3 and below	1.	0	6	25	19	29	1	80
		0.0	7.5	31.2	23.7	36.2	1.2	21.6
		0.0	33.3	43.1	23.5	15.3	4.2	
		0.0	1.6	6.7	5.1	7.8	0.3	
11.4 to 13.9	2.	1	8	16	25	43	2	95
		1.1	8.4	16.8	26.3	45.3	2.1	25.6
		100.0	44.4	27.6	30.9	22.8	8.2	
		0.3	2.2	4.3	6.7	11.6	0.5	
14.0 and above	3.	0	4	17	37	117	21	196
		0.0	2.0	8.7	18.9	59.7	10.7	52.8
		0.0	22.2	29.3	45.7	61.9	87.5	
		0.0	1.1	4.6	10.0	31.5	5.7	
	Column Total	1 / 0.3	18 / 4.9	58 / 15.6	81 / 21.8	189 / 50.9	24 / 6.5	371 / 100.0

Chi square = 48.74638 with 10 degrees of freedom
Significance = 0.0000

Table 8.7.5 Cross-tabulation: EPTB (GRDT) by tutors' ratings (RT)

Count / Row PCT / Col PCT / Tot PCT	RT 1 Inadequate	2 Borderline	3 Adequate	4 Completely Adequate	Row Total
GRDT 1. 33.9 and below	1	16	26	7	50
	2.0	32.0	52.0	14.0	17.2
	16.7	36.4	13.8	13.2	
	0.3	5.5	8.9	2.4	
2. 34.0 to 39.9	4	24	102	15	145
	2.8	16.6	70.3	10.3	49.8
	66.7	54.5	54.3	28.3	
	1.4	8.2	35.1	5.2	
3. 40.0 to 45.9	1	4	49	22	76
	1.3	5.3	64.5	28.9	26.1
	16.7	9.1	26.1	41.5	
	0.3	1.4	16.8	7.6	
4. 46.0 and above	0	0	11	9	30
	0.0	0.0	55.0	45.0	6.9
	0.0	0.0	5.9	17.0	
	0.0	0.0	3.8	3.1	
Column Total	6	44	188	53	291
	2.1	15.1	64.6	18.2	100.0

Chi square = 38.84460 with 9 degrees of freedom
Significance = 0.0000

Table 8.7.6 Cross-tabulation: BCSA (GRST) by tutors ratings (RT)

Count / Row PCT / Col PCT / Tot PCT		RT 1 Inadequate	2 Borderline	3 Adequate	4 Completely Adequate	Row Total
GRST 11.3 and below	1.	1 1.3 50.0 0.3	11 13.9 33.3 3.1	60 75.9 27.0 17.0	7 8.9 7.4 2.0	79 22.4
11.4 to 13.9	2.	1 1.2 50.0 0.3	11 13.1 33.3 3.1	56 66.7 25.2 15.9	16 19.0 16.8 4.5	84 23.9
14.0 and above	3.	0 0.0 0.0 0.0	11 5.8 33.3 3.1	106 56.1 47.7 30.1	72 38.1 75.8 20.5	189 53.7
Column Total		2 0.6	33 9.4	222 63.1	95 27.0	352 100.0

Chi square = 31.78748 with 6 degrees of freedom
Significance = 0.0000

Section 3 CPM sample data

Table 8.8 CPM sub-samples: distribution by subject area

Subject Area	EPTB Freq %	BCSA Freq %	CPM Freq%	Whole sample Freq %
1. Agriculture, etc	2.9	5.4	4.2	8.4
2. Arts	-	-	-	2.9
3. Medicine, etc	2.9	7.1	5.3	7.3
4. Pure sciences	2.9	12.5	9.5	11.4
5. Education	32.4	39.3	34.7	16.2
6. Engineering, etc	11.8	1.8	9.5	21.8
7. English studies, etc	2.9	3.6		
8. Social sciences (prof)	44.1	26.8	31.6	18.0
9. Social sciences (acad)	-	3.6	2.1	7.3
10. Miscellaneous	-	-	-	0.5
N	34	56	95	100.0

Note: EPTB and BCSA columns do not include cases for whom tutors' ratings were not obtained.

Table 8.9 CPM sub-samples: distribution by level of study

Subject Area	EPTB Freq %	BCSA Freq %	CPM Freq%	sample Freq %
2. Pre-university	-	-	-	1.2
3. First degree	-	-	-	5.5
4. Masters by tuition	14.7	10.7	16.8	23.3
5. Research (degree)	2.9	7.1	5.3	16.0
6. Professional diploma	82.4	82.1	77.9	41.6
7. Academic attachment	-	-	-	9.5
8. Professional attachment	-	-	-	2.9
	34	56	95	100.0

Note: EPTB and BCSA columns do not include cases for whom tutors' ratings were not obtained.

Table 8.10 CPM sub-samples: distribution of EPTB and BCSA results

	EPTB		Freq	Freq %	Whole sample Freq %
1.	less than 34.0	English inadequate	5	14.7	16.7
2.	34.0 to 39.9	English tuition needed	25	73.5	50.8
3.	40.0 to 45.9	English adequate	4	11.8	26.0
4.	46.0 and above	Unquestionably adequate	-	-	6.4
		Total	34	100.0	100.0
		Mean	37.62		38.32
	BCSA				
1.	C+ and below	English probably inadequate	9	16.1	21.6
2.	B, B+ (low)	English tuition needed	11	19.6	25.6
3.	B+ (high), A	English adequate	36	64.3	52.8
		Total	56	100.0	100.0
		Mean	low B+ (3.43)		low B+ (3.34)

Table 8.11 CPM sub-samples: length of remedial English tuition

No of weeks	EPTB Freq %	BCSA Freq %	CPM Freq %	Whole sample frequency %
Up to 4	18.5	28.9	24.3	30.4
5 to 8	40.7	34.2	35.7	31.1
9 to 12	40.7	34.2	38.6	32.8
13 or more	-	2.6	1.4	5.7
Total	34	56	100.0	100.0
None, as % of sample	20.6	32.1	26.3	36.9

Section 4 Pre-departure assessments and the criterion measures

Table 8.12 Correlations: EPTB with CPM

		CPM				
		C	W	S	WS	O
EPTB	D1	0.2187 (32) P=0.115	0.0472 (31) P=0.400	0.1134 (30) P=0.275	0.0767 (29) P=0.346	0.2163 (30) P=0.126
	D2	0.1134 (32) P=0.268	0.0312 (31) P=0.434	0.1451 (30) P=0.222	0.1006 (29) P=0.302	-0.0314 (30) P=0.345
	D3	0.2545 (33) P=0.076	0.3300 (32) P=0.033	0.2690 (31) P=0.072	0.3360 (30) P=0.035	0.2829 (31) P=0.062
	D4	0.3849 (33) P=0.013	0.3171 (32) P=0.039	0.2641 (31) P=0.076	0.3339 (30) P=0.035	0.4242 (31) P=0.009
	DT	0.4363 (34) P=0.005	0.3256 (3) P=0.032	0.3144 (32) P=0.040	0.3298 (31) P=0.035	0.3356 (32) P=0.030
	D5	0.7567 (13) P=0.001	0.8511 (13) P=0.000	0.7415 (12) P=0.003	0.8571 (12) P=0.000	0.6924 (12) P=0.006

Key:
- D1 = Phonemic discrimination
- D2 = Intonation
- D3 = Reading Comprehension
- D4 = Grammar
- DT = EPTB Pt 1 Total
- D5 = EPTB Pt 2 (reading speed)

- C = Total cloze test score
- W = Writing level
- S = Interview level
- WS = Composite Writing/Interview level
- O = Overall CPM level

Table 8.13 Correlations: BCSA with CPM

		CPM				
		C	W	S	WS	O
BCSA	D1	0.2885 (45) P=0.027	0.1864 (43) P=0.116	0.2801 (38) P=0.044	0.2294 (36) P=0.089	0.1609 (38) P=0.167
	D2	0.4010 (45) P=0.003	0.5192 (43) P=0.000	0.6207 (38) P=0.000	0.6787 (36) P=0.000	0.5433 (38) P=0.000
	D3	0.4905 (44) P=0.000	0.3372 (42) P=0.014	0.4279 (37) P=0.004	0.5022 (35) P=0.001	0.4138 (37) P=0.005
	D4	0.4415 (44) P=0.001	0.4907 (42) P=0.000	0.4836 (37) P=0.001	0.5776 (35) P=0.000	0.5354 (37) P=0.000
	DT	0.5029 (55) P=0.000	0.5294 (53) P=0.000	0.5602 (48) P=0.000	0.6619 (45) P=0.000	0.5676 (47) P=0.000

Key:
- D1 = Listening
- D2 = Speaking
- D3 = Reading
- D4 = Writing
- DT = Total Grades

- C = Total cloze test score
- W = Writing level
- S = Interview level
- WS = Composite Writing/Interview level
- O = Overall CPM level

Table 8.14 Correlations (CPM sub-sample): EPTB with tutors' ratings

		Tutors' Ratings				
		RL	RS	RR	RW	RT
EPTB	D1	0.0074	-0.0688	0.1391	-0.1559	-0.0391
		(32)	(32)	(30)	(30)	(30)
		P=0.484	P=0.354	P=0.232	P=0.205	P=0.419
	D2	-0.0758	0.0108	-0.0449	0.0435	-0.0150
		(32)	(32)	(30)	(30)	(30)
		P=0.340	P=0.477	P=0.407	P=0.410	P=0.469
	D3	0.2289	0.1883	0.4922	0.1186	0.2691
		(33)	(33)	(31)	(31)	(31)
		P=0.100	P=0.147	P=0.002	P=0.263	P=0.072
	D4	-0.-345	0.1869	-0.0453	0.3006	0.1339
		(33)	(33)	(31)	(31)	(31)
		P=0.424	P=0.149	P=0.404	P=0.050	P=0.236
	DT	0.0446	0.1571	0.2190	0.1602	0.1589
		(34)	(34)	(32)	(32)	(32)
		P=0.401	P=0.187	P=0.114	P=0.191	P=0.193
	D5	0.4678	0.6710	0.5960	0.7614	0.6960
		(13)	(13)	(13)	(13)	(13)
		P=0.053	P=0.006	P=0.016	P=0.001	P=0.004

Key: D1 = Phonemic discrimination
D2 = Intonation
D3 = Reading Comprehension
D4 = Grammar
DT = EPTB Pt 1 Total
D5 = EPTB Pt 2 (reading speed)

Tutors' Ratings
RL = Listening
RS = Speaking
RR = Reading
RW = Writing
RT Total Skills

Table 8.15 Correlations (CPM sub-sample): BCSA with tutors' ratings

		Tutors' Ratings				
		RL	RS	RR	RW	RT
BCSA	D1	0.1685 (45) P=0.134	0.0933 (45) P=0.271	0.1662 (42) P=0.146	0.0929 (40) P=0.284	0.1616 (40) P=0.160
	D2	0.2819 (45) P=0.030	0.3881 (45) P=0.004	0.2101 (42) P=0.091	0.1869 (40) P=0.124	0.2814 (40) P=0.039
	D3	0.2203 (44) P=0.075	0.1326 (44) P=0.195	0.1834 (41) P=0.126	0.1533 (39) P=0.176	0.1982 (39) P=0.113
	D4	0.2487 (44) P=0.052	0.3037 (44) P=0.023	0.1458 (41) P=0.182	0.3292 (39) P=0.020	0.2946 (39) P=0.034
	DT	0.2848 (56) P=0.017	0.3223 (56) P=0.008	0.2435 (53) P=0.039	0.2041 (51) P=0.075	0.2967 (51) P=0.017

Key: D1 = Listening
D2 = Speaking
D3 = Reading
D4 = Writing
DT = Total Grades

Tutors' Ratings
RL = Listening
RS = Speaking
RR = Reading
RW = Writing
RT Total Skills

Table 8.16 Contingency tables - EPTB with CPM

CPM (overall)

EPTB	1.	2.
1.	0*	5
2.	2	28*

Expectations confirmed = 80%
Source: Chi square = 21.6
df = 6, P = .001

CPM (oral/writing)

EPTB	1.	2.
1.	0*	5
2.	2	27*

Expectations confirmed = 79.5%
Source: Chi square = 32.4
df = 16, P = .009

Table 8.17 Contingency tables - EPTB with tutors' ratings (CPM sample)

Ratings

EPTB	1.	2.
1.	0*	5
2.	3	24*

Expectations confirmed = 75%
Source: Chi square = 15.2
df = 6, P = .018

Table 8.18 Contingency tables - BCSA with CPM

CPM (overall)

BCSA	1.	2.
1.	0*	7
2.	3	39*

Expectations confirmed = 79.5%
Source: Chi square = 16.3
df = 6, P = .012

CPM (oral/writing)

BCSA	1.	2.
1.	0*	6
2.	2	38*

Expectations confirmed = 82.5%
Source: Chi square = 28.6
df = 16, P = .026

Table 8.19 Contingency tables - BCSA with tutors' ratings (CPM sample)

Ratings

BCSA	1.	2.
1.	0*	9
2.	0	42*

Expectations confirmed = 82%
Source: Chi square = 17.1
df = 4, P = .002

Section 5 Pre-departure measures and tutors' ratings, according to different variables

Table 8.20 Data - sub-samples by sex and remedial English

(distributions in %)

	Whole Sample	Female	Male	+Rem	-Rem
N	729	117	612	460	269
Age - mean years	30.3	30	30.3	30.6	29.7
Education - secondary	20.6	18.8	20.9	25.4	12.3
- first degree	59.7	62.4	59.1	60.2	58.7
- master's degree	16.0	16.2	16.0	11.1	24.5
- research degree	3.4	2.6	3.6	2.8	4.5
- other	0.3	-	0.3	0.4	-
Subject - agriculture	8.4	3.4	9.3	9.2	7.1
- arts	2.9	4.3	2.6	2.4	3.7
- medicine	7.3	10.3	6.7	6.1	9.3
- pure sciences	11.4	12.8	11.1	10.0	13.8
- education	16.2	28.2	13.9	16.8	15.2
- engineering	21.8	5.1	25.1	22.0	21.6
- English etc	6.2	12.8	4.9	4.1	9.7
- social sci prof	18.0	17.1	18.2	21.8	11.5
- social sci acad	7.3	6.0	7.4	6.8	8.2
- miscellaneous	0.5	-	0.7	0.9	-
Level - first degree	5.5	4.3	5.7	5.7	5.2
- master's degree	23.3	15.4	24.9	22.6	24.5
- research degree	16.0	17.1	15.7	10.0	26.4
- diploma	41.6	50.4	39.9	44.8	36.1
- acad attachment	9.5	12.0	9.0	11.7	5.6
- prof attachment	2.9	0.9	3.3	4.1	0.7
EPTB - less than 34	16.7	15.0	17.0	23.3	3.8
- 34.0 to 39.9	50.8	42.5	52.0	61.2	30.5
- 40.0 to 45.9	26.0	32.5	25.1	15.0	47.6
- 46 and over	6.4	10.0	5.9	0.5	18.1
mean score	38.3	39.0	38.2	36.6	41.5
BCSA - C+ or less	21.6	17.1	22.7	32.9	2.2
- B-, B	25.6	31.4	24.3	35.5	8.8
- B+, A	52.8	51.4	53.0	31.6	89.1
mean score	B+	13.6 (B+)	13.3 (B+)	12.3 (B)	15.1 (A-)
Remedial English	63.1	60.7	63.7	-	-
Ratings - inadequate	2.2	2.6	2.1	3.5	-
(R1A) - only just adeq	21.6	17.9	27.0	12.4	12.1
- adequate	51.2	47.9	53.6	47.2	47.2
- completely adeq	24.9	31.6	15.9	40.4	40.4
Mean skills rating (RT)	9.3	9.4	9.3	8.8	10.1
Improvement - considerable	20.2	13.7	21.5	22.8	15.7
- some	56.0	53.8	56.3	61.3	46.8
- n/a	12.4	15.4	11.8	8.0	19.9
Correlation: EPTB/R1A	.31	NS	.32	.13	.42
EPTB/RT	.36	.39	.35	.19	.42
BCSA/R1A	.36	.44	.35	.25	.06*
BCSA/RT	.36	.46	.35	.23	-.04*

Table 8.21 Contingency tables: EPTB/BCSA by groups according to sex

(a) EPTB group (female)

EPTB	Rating (R1A) 1.	2.
1.	0	6
2.	0	34

Expectations confirmed = 85%
Source: Chi square = 4.88
df = 6, P = .559

(b) BCSA group (female)

BCSA	Rating (R1A) 1.	2.
1.	1	11
2.	2	56

Expectations confirmed = 81%
Source: Chi square = 15.4
df = 6, P = .01

(c) EPTB group (male)

EPTB	Rating (R1A) 1.	2.
1.	3	43
2.	6	219

Expectations confirmed = 81%
Source: Chi square = 40.4
df = 9, P = 00

(d) BCSA group (male)

BCSA	Rating (R1A) 1.	2.
1.	2	66
2.	2	227

Expectations confirmed = 77%
Source: Chi square = 37.5
df = 6, P = .00

Table 8.22 Means and SD's for the remedial English groups

			Pre-departure	Rating (RT)
EPTB Pt 1	Total sub-sample	M	38.3	35.6
		SD	4.5	7.6
	No remedial group	M	40.6	38.3
	Plus remedial group	M	35.3	34.2
	Difference		over 1 SD	over ½ SD
BCSA total	Total sub-sample	M	13.3 (B+)	38.6
		SD	2.2	7.1
	No remedial group	M	15.1 (A-)	42.1
	Plus remedial group	M	12.3 (B)	36.3
	Difference		1½ SD	over ¾ SD

Table 8.23 Contingency tables: EPTB/BCSA by groups according to remedial English

(a) EPTB by + Remedial English

		Rating (R1A) 1.	2.
EPTB	1.	3	45
	2.	6	152

Expectations confirmed = 75%
Source: Chi square = 12.6
df = 9, P = .18

(b) BCSA by + Remedial English

		Rating (R1A) 1.	2.
BCSA	1.	3	74
	2.	4	152

Expectations confirmed = 66.5%
Source: Chi square = 23.1
df = 6, P = .00

(c) EPTB by No Remedial English

		Rating (R1A) 1.	2.
EPTB	1.	0	4
	2.	0	101

Expectations confirmed = 96%
Source: Chi square = 24.6
df = 6, P = .00

(d) BCSA by No Remedial English

		Rating (R1A) 1.	2.
BCSA	1.	0	3
	2.	0	132

Expectations confirmed = 97.5
Source: Chi square = 5.3
df = 4, P = .25

Table 8.24.1 Correlations: EPTB with Remedial English groups

(a) with + Remedial English group

	DT
R1	0.1336 (206) P=0.028
R2	0.1427 (206) P=0.020
RT	0.1890 (191) P=0.004

(b) with No Remedial English group

	DT
R1	0.4196 (105) P=0.000
R2	0.2312 (105) P=0.009
R3	0.4248 (100) P=0.000

Key: DT = EPTB Pt 1 Total score
R1 = Tutors' ratings - question 1A
R2 = Tutors' ratings - question 1B
RT = Tutors' total skill ratings - question 2

Table 8.24.2 Correlations: BCSA with Remedial English groups

(a) with + Remedial English group

	DT
R1	0.2541 (233) P=0.000
R2	0.2228 (234) P=0.000
RT	0.2317 (222) P=0.000

(b) with No Remedial English group

	DT
R1	0.0555 (135) P=0.261
R2	0.0046 (137) P=0.479
RT	-0.0357 (130) P=0.343

Key: DT = BCSA total grade
R1 = Tutors' ratings - question 1A
R2 = Tutors' ratings - question 1B
RT = Tutors' total skills ratings - question 2

Table 8.26 Correlations: EPTB and BCSA with tutors' ratings by subject area

Subject area	N	Age	DT	RT	r DT/ R1A	r DT/ RT	CE %
EPTB sub-sample	316	30.3	38.3	8.9	.31	.36	82
Agricultural, veterinary	25	30.8	36.8	8.5	.20*	.35	84
Arts	9	30.3	41.3	10.1	NS	NS	-
Medicine, nursing etc	19	32	38.2	9.2	.05*	.23*	79
Physical, biological sciences	34	29.1	38.7	8.8	.15*	.24*	82*
Education, TEFL	39	34.5	36.5	8.8	.52	.60	66
Engineering, technology	84	27.4	39.3	9.1	.35	.40	86
English studies etc	10	34.5	43.4	10.3	.05*	.18*	100
Business, social sciences (professional)	66	30.8	37.6	8.4	.16*	.22	80
Social sciences (academic)	27	27.6	38.4	8.8	.42	.31*	92
BCSA sub-sample	413	30.2	13.3(B+)	9.6	.36	.36	78
Agricultural, veterinary	36	30.2	13.5(B+)	10.3	-.02*	.00*	82
Arts	14	27.8	13.3(B+)	10.7	.30*	.36*	-
Medicine, nursing etc	34	31.7	13.4(B+)	9.8	.35	.40	82
Physical, biological sciences	48	30.1	13.1(B+)	10.1	.62	.50	78
Education, TEFL	79	32.7	13.5(B+)	9.8	.22	.23	83
Engineering, technology	75	28.2	13.0(B+)	9.3	.43	.49	73
English studies etc	35	31.8	13.9(B+)	9.1	.60	.54	86
Business, social sciences (professional)	65	29.7	13.5(B+)	9.1	.43	.48	79
Social sciences (academic)	26	28.6	13.6(B+)	9.7	.56	.37	74

Table 8.27.1 EPTB and BCSA distributions and comparisons with tutors' ratings by level of study

	First Degree	Master's Degree	Research (Degree)	Diploma	Acadamic Attachment
EPTB group					
N	17	88	45	115	28
EPTB scores:					
less than 34.0	-	12%	16%	14%	46%
34.0 to 39.9	23%	50%	44%	60%	36%
40.0 to 45.9	41%	33%	36%	20%	14%
46 or more	35%	4%	4%	6%	4%
Correlations:					
DT/R1	.28*	.35	.25	.32	.19*
DT/RT	.19*	.35	.43	.38	.36
Confirmed expectancy rate	100%	88%	84%	83%	54%
BCSA group					
N	23	82	72	188	41
BCSA grades:					
C+ or less	31%	27%	10%	20%	24%
B-, B	15%	25%	22%	26%	37%
B+, A	54%	48%	67%	55%	39%
Correlations:					
DT/R1	-.22*	.35	.57	.32	.13*
DT/RT	-.06*	.34	.58	.31	.31
Confirmed expectancy rate	69%	70%	90%	81%	74%
Tutors' total skills rating (RT)	9.3	9.2	10.1	9.2	9.4
Mean age	23.6	29.1	30.5	31.2	32.7

* denotes not significant at the 5% level